Design, evaluation and application of methodology and software for time-to-event outcomes in pharmacogenetic genome-wide association studies

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Hamzah Syed

May 2018

UNIVERSITY OF LIVERPOOL

ABSTRACT

Thesis title: Design, evaluation and application of methodology and software for time-to-event outcomes in pharmacogenetic genome-wide association studies
Author: Hamzah Syed
Introduction and aims: Methodology and software for the analysis of genome-wide association studies (GWAS) have focused on binary phenotypes and quantitative traits. However, the impact of single nucleotide polymorphisms (SNPs) on time-to-event (TTE) outcomes is understudied, particularly for pharmacogenetic GWAS. Statistical methodology and computational tools to design and analyse GWAS with TTE outcomes are not well developed, due to the scale and complexity of data, particularly when analysing rare variants. This thesis aims to develop statistical methodology and a variety of computational tools to aid the design and analysis of both GWAS of common and rare variants with TTE outcomes.
Methods: This thesis compares existing methodology such as the Cox proportional hazards, logistic and Weibull regression models using simulations based on a range of pharmacogenetic GWAS designs with TTE outcomes. This thesis also presents new statistical methodologies for the analysis of rare variants using a combination of gene-based tests of association and TTE regression models.
Results: Examination of the literature provided an overview of the methods and software used for analysing GWAS with TTE outcomes. One approach taken due to lack of software availability was to dichotomise event times at a fixed time-point and analyse the binary outcome using existing GWAS software. A simulation study was conducted comparing alternative regression models under pharmacogenetic TTE study designs. This simulation study demonstrated that dichotomisation of the TTE outcome would result in a loss of statistical power. Hence, the thesis outlines three user-friendly computational tools specific to TTE GWAS. The first is SurvivalGWAS_Power, which performs power calculations and generates sample pharmacogenetic data across a range of design scenarios, allowing for censoring and interactions. Second, SurvivalGWAS_-SV, software capable of analysing large-scale imputed GWAS data, offering a variety of survival analysis models. Third, rareSurvival, a command line application, which implements gene-based burden tests for the analysis of rare variants with TTE outcomes. SurvivalGWAS_SV and rareSurvival have been evaluated through simulation studies as well as application to a GWAS investigating the pharmacogenetics of acute coronary syndrome (PhACS). The single variant and gene discovery analyses of the PhACS study identified novel loci associated with time to recurrence of a cardiovascular event including rs56045815 located in the *CTNNA2* gene.
Conclusions: This thesis introduces three novel computational tools for GWAS with TTE outcomes. SurvivalGWAS_SV and rareSurvival are compatible with high-performance computing clusters and are available on Linux, Windows and Mac OSX operating systems. SurvivalGWAS_SV and rareSurvival were applied to the PhACS data, identifying significantly associated SNPs and functional units for further follow-up. With their particular relevance to pharmacogenetic GWAS, SurvivalGWAS_Power, SurvivalGWAS_SV and rareSurvival, will help in the design of studies and identification of genetic biomarkers of patient response to treatment, with the ultimate goal of personalising therapeutic interventions.

ACKNOWLEDGEMENTS

The past three and a half years have been the most exciting and influential chapter of my life. The experiences I have gained, the friends and family I have made along the way are what have shaped me into becoming a better person.

I would first like to thank my supervisors Professor Andrew Morris and Dr Andrea Jorgensen for giving me this opportunity and investing a lot of time into my development as an academic. Thanks to their generosity and guidance, I have been able to achieve my full potential. For this, I am eternally grateful to them. I would like to extend my thanks to the Department of Biostatistics at the University of Liverpool for funding my research. With the funding I received, I was able to publish and present my work at conferences all over the world. To see the impact of my research on a global scale was rewarding.

I am very grateful for my friends and colleagues who have helped me in numerous ways over the years, especially for Dr James Cook and Dr Ben Francis. They have been my mentors and two of my closest friends during my PhD. I will always cherish the burrito and doughnut lunches we shared on those particularly stressful days.

I am thankful to my father, Kalimullah Syed for always encouraging me to give 110% in everything that I do, and my mother, Dr Asfia Syed whose academic accomplishments have truly inspired me. Without their support, I would not have taken on this challenge.

My final thanks go to the person who is most important to me and to whom I dedicate this thesis to, my wife and best friend, Tuğçe. Not only for being there for me but for carrying me through this journey. We both pushed each other to be better, studying side by side into the dead of night. She took care of me through the months after my heart surgery and kept me sane on those days when the thesis became overwhelming for me. Our unique journey took detours through many life-changing moments, and I could not imagine doing any of it without her.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SCRIPTS

STATEMENT OF DATA CONTRIBUTION

The access and use of the Standard And New Anti-epileptic Drug (SANAD) trial data was approved by Professor Tony Marson (Principal Investigator). The Health Technology Assessment funded the original trial. The views and opinions expressed within this thesis do not necessarily reflect those of the National Health Service, the Health Technology Assessment or the Department of Health.

Chapter 5 describes the use of exome array genotype data from the Prospective Investigation of Vasculature in Uppsala Seniors (PIVUS) and the Uppsala Longitudinal Study of Adult Men (ULSAM). Only the genotype data were used for simulations, and no phenotype information was made available. Use of these data were approved by Professor Andrew Morris (Principal Investigator). The use of the Pharmacogenetics of Acute Coronary Syndrome (PhACS) data outlined in Chapter 6, was approved by Professor Sir Munir Pirmohamed (Principal Investigator). The PhACS study was funded by the UK Department of Health as part of the NHS Chair in Pharmacogenetics Programme.

PUBLISHED CONTENT

Syed, H., Jorgensen, A. L. & Morris, A. P. 2016a. Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies. Pharmacogenomics, 17, 907-15.

Syed, H., Jorgensen, A. L. & Morris, A. P. 2016b. SurvivalGWAS_Power: a user friendly tool for power calculations in pharmacogenetic studies with "time to event" outcomes. BMC Bioinformatics, 17, 523.

Syed, H., Jorgensen, A. L. & Morris, A. P. 2017. SurvivalGWAS_SV: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes. BMC Bioinformatics, 18, 265.

ABBREVIATIONS

A/D	Hospital Admission
ACEI	Angiotensin-Converting Enzyme Inhibitor
AF	Acceleration Factor
AFT	Accelerated Failure Time
API	Application Programming Interface
BT	Burden Test
BMI	Body Mass Index
CABG	Coronary Artery Bypass Grafting
CEU	Utah residents with Northern and Western European ancestry
CHB	Han Chinese in Beijing
CHR	Chromosome
CPHM	Cox Proportional Hazards Model
CVD	Cardiovascular Disease
DNA	Deoxyribonucleic Acid
D/C	Hospital Discharge
EAF	Effect Allele Frequency
EOS	End Of Study
FAQ	Frequently Asked Questions
GLM	Generalised Linear Model
GPC	Genetic Power Calculator
GPL	General Public License
GUI	Graphical User Interface
GWAS	Genome-Wide Association Study

HPC High Performance Computing

HR Hazard Ratio

HWE Hardy-Weinberg Equilibrium

IBD Identity By Descent

IBS Identity By State

JPT Japanese in Tokyo

LCI Lower Confidence Interval

LD Linkage Disequilibrium

LRT Likelihood Ratio Test

O/S Operating System

OS Overall Survival

MAF Minor Allele Frequency

MI Myocardial Infarction

PH Proportional Hazards

PhACS Pharmacogenetics of Acute Coronary Syndrome

PC Principal Component

PCA Principal Component Analysis

PCI Percutaneous Coronary Intervention

QQ Quantile-Quantile

RAM Random Access Memory

RVAS Rare Variant Association Study

SANAD Standard And New Anti-epileptic Drug

SKAT Sequence Kernel Association Test

SNP Single Nucleotide Polymorphism

TTE Time To Event

UCI Upper Confidence Interval

UK United Kingdom

VAT Variant Association Tool

VCF Variant Call Format

WTCCC Wellcome Trust Case Control Consortium

YRI Yoruba in Ibadan

CHAPTER 1

INTRODUCTION

In the last decade, genome-wide association studies (GWAS) have become the traditional approach for the discovery of genetic variations contributing to a multitude of complex human traits and diseases. GWAS aim to test the association between genetic markers across the genome with a particular phenotype[1] within a population of (typically) unrelated individuals. Prior to this, candidate gene studies were undertaken based on functional studies and existing biological/clinical knowledge of the trait. A candidate gene study aims to thoroughly examine a gene, studying variations that are both common and rare in a population. However, this approach relies on prior knowledge, with many studies reporting failure to replicate the causal genes previously found. GWAS differ from candidate gene studies in some important ways: (i) GWAS are discovery-driven as opposed to hypothesis-driven; and (ii) a greater number of variants are investigated through GWAS.

The GWAS era rose to popularity in the mid-2000s, consequently spawning the publication of the landmark GWAS study of 14,000 cases of seven common diseases and 3,000 shared controls conducted by the Wellcome Trust Case Control Consortium (2007) (WTCCC). This study was one of the first large-scale GWAS to be undertaken. Following this, in the last decade, thousands of GWAS have been published with associations catalogued in the NHGRI-EBI Catalog (MacArthur et al. 2017). GWAS have produced an extraordinary amount of discoveries of genomic regions associated with disease risk and many biological characteristics.

The current chapter provides an introduction to the basic concepts of genetic research, with a focus on GWAS, from inception to present day and with a brief description of the use of GWAS within the field of pharmacogenetics. Explanations are given on

[1] An observable characteristic or trait expressed by the genotype.

the design of GWAS, quality control procedures and association testing. Methodology for survival analysis is discussed in the context of genetic association studies, with particular attention given to defining the different models used throughout this thesis, including phenotype definitions and the incorporation of genetic modes of inheritance within regression models. Following this, the role of computational tools is examined, highlighting the latest software innovations. Finally, this introduction concludes with an overview of all the chapters, defining the aims and explaining the motivation behind the research in this thesis.

1.1 Background to Genetics Research in Healthcare

One of the major goals of genetic research is to improve healthcare through the prediction of future disease and personalisation of treatments with therapeutic benefits for an individual. One of the building blocks for attaining this goal is the discovery of genetic biomarkers[2] (i.e. genetic variants) associated with a disease or treatment response. Methodology and software for analysis are at the forefront of this discovery process.

1.1.1 The Genome and Single Nucleotide Polymorphisms

Genetics is defined as the study of heredity, specifically, transmission of characteristics from one generation to the next (Teare 2011). Nuclei of human cells store the genetic information coded by deoxyribonucleic acid (DNA). The genome is a complete set of DNA made up of a chain of linked nucleotide bases, adenine (A), thymine (T), cytosine (C) and guanine (G). Bases are nitrogen-containing biological compounds and are the building blocks for nucleic acids. There are a total number of 22 pairs and two sex chromosomes, along with 30,000 genes contained in the human genome. Genes are specific sequences of bases at particular loci that encode instructions on how to make proteins[3]. A locus is a fixed position on a chromosome or region of the

[2]A biomarker or biological marker is a measurable predictor of a biological state, treatment response or presence of a disease. In genetics research, this can be a measurable characteristic, SNP or gene.
[3]Proteins determine the function of a cell.

genome. This fixed position can represent a gene or any interval of variants. Most loci are identical between individuals, however, several different types of variations exist such as copy number variants[4] and indels[5]. The most common types of variations in the DNA sequence are single nucleotide polymorphisms (SNPs), which occur in at least 1% of the population and represent a single base change. The 'value' of these variants is called the genotype. These genotypes are the different allelic combinations that can be present at a SNP. For example, consider a bi-allelic variant, meaning that there are only two alleles in a specific locus, where A is the major (most frequent) allele, and C is the minor (least frequent) allele. The genotypes are represented as AA (major homozygous), AC (heterozygous) and CC (minor homozygous). The focus of genetic association analysis is centred around SNPs as genetic biomarkers. There are over 10 million SNPs within the human genome. Many are important as they underlie differences in our traits, such as eye colour and our susceptibility to disease. However, there is still much unknown about the function of the majority of SNPs in the genome. These variants are routinely analysed due to their high density in the genome and the fact that they are relatively easy to genotype using high throughput technology.

1.2 Genome-Wide Association Studies

The objective of GWAS is to identify SNPs that are associated with a trait, covering millions of variants across the genome in samples of individuals within a study population. The GWAS approach enables the detection of variants with varying sizes of effect on phenotype amongst a vast amount of variants, utilising a stringent threshold for statistical significance $\left(p < 5 \times 10^{-8}\right)$ to reduce the number of false positive associations due to multiple testing. This significance threshold is adapted from the Bonferroni correction (Bland & Altman 1995) for the number of independent statistical tests performed. However, we expect that SNPs across the genome are correlated. Therefore the 5×10^{-8} threshold is widely-accepted as genome-wide significance based on ap-

[4]Copy number variants or CNVs are a section of variants that are duplicated across the DNA sequence.
[5]Indels are insertions or deletions of a range of base-pairs across the DNA sequence.

proximately 1 million blocks of linkage disequilibrium (LD) (see Section 1.2.1) across the genome (Pe'er et al. 2008). A variety of corrections for multiple testing exist with the debate still continuing on the optimal threshold (Panagiotou et al. 2012, Kanai et al. 2016, Fadista et al. 2016) especially with the increasing availability of whole-genome sequence data where many more variants are interrogated than in GWAS (see Section 1.2.2).

GWAS analysis techniques typically assume that each variant being investigated is equally likely to be associated with the outcome of interest in an unbiased way. Doing this maximises the opportunity for the discovery of variants not known, a priori, to be biologically relevant to the trait under investigation.

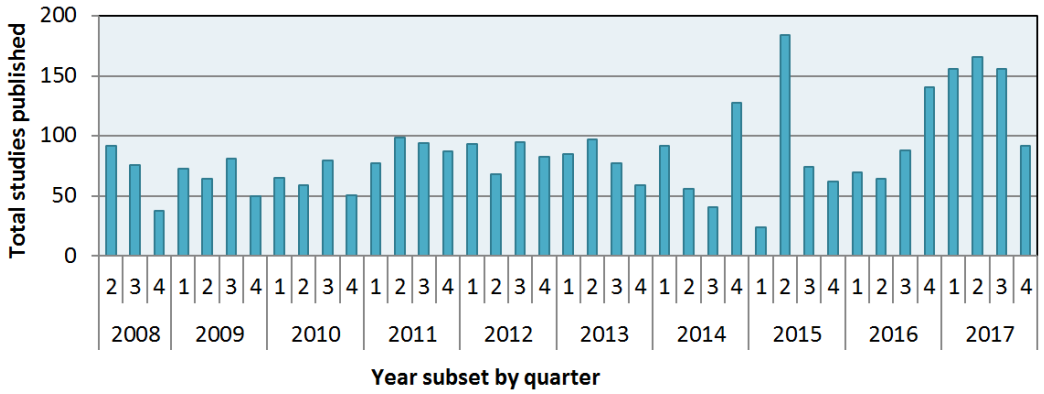


Figure 1.1: Historical data of published GWAS from the NHGRI-EBI Catalog. Data is from the second quarter of 2008 to the fourth quarter of 2017.

1.2.1 Allele Frequency and Linkage Disequilibrium

Determining the allele frequency of a SNP is key to defining the genetic diversity within a study population. The major allele frequency, p, is defined as the frequency of the allele which occurs most commonly within a population. The minor allele frequency (MAF) is thus $1 - p$. Where alleles are independent of one another, the loci are said to be in Hardy-Weinberg equilibrium (HWE). This theorem is based on the assumption that in the absence of any evolutionary influences such as selection and assortative mating the allele and genotype frequencies in a population will remain constant from

one generation to the next. Under this condition the genotype frequencies are derived as p^2, $2p(1-p)$ and $(1-p)^2$ for the major homozygous, heterozygous and minor homozygous genotypes respectively. Testing SNPs for deviation from HWE is covered in Section 1.2.3.

The international HapMap Project (Altshuler et al. 2010) showed that SNPs are arranged in blocks of strong linkage disequilibrium (LD). LD is an important term used when an allelic correlation is identified between two or more loci within a population, suggesting that genotypes at one locus are correlated with genotypes at a second locus. LD is important for the discovery and localisation of genes (gene-mapping) associated with a trait to reveal new insights. The strength of the relationship between loci can be quantified using many different metrics. For instance, the Pearson's correlation of alleles is defined as r^2, which is a value between 0 and 1, where 1 signifies perfect correlation between alleles at one SNP and a second SNP. Many other measures exist, and Teare (2011) explains them in detail, providing formulae and descriptions on the benefits of each metric. LD between SNPs across a locus can be visualised and explored further using the web-based tool LDlink (Machiela & Chanock 2015).

1.2.2 Genotyping

With the advancement of high throughput technologies for genotyping using genome-wide chip arrays, came a surge in the number of GWAS undertaken for common variation (see Figure 1.1), typically defined as variants with a MAF of at least 1%. Knowledge of the LD patterns across the genome has helped in the design of genotyping products. Often the variants from genotype arrays are supplemented via imputation to increase coverage[6]. Imputation seeks to make inferences about the genotype of untyped individuals for a group of variants based on the known phased haplotypes[7] from densely genotyped individuals. Examples of such high-density reference data for imputation

[6]Coverage is how much of the human genome can be inferred from the genotype chip array. Coverage in the context of sequencing is the number of sequence reads that have alignments that overlap a certain position. The greater the coverage, the fewer the sequencing errors in base calling.

[7]A haplotype is a group of specific alleles for different SNPs on a single chromosome.

include initiatives such as the 1000 Genomes Project (Auton et al. 2015). The project sequenced the whole genomes of more than 2500 individuals from 26 populations, undertaken to increase the number of individuals and populations represented and to extend variant coverage to lower MAFs.

The introduction of next-generation sequencing (NGS) capabilities such as whole-genome and -exome sequencing provide complete coverage of variation in individuals, even lower frequency variants. This process is more costly than array genotyping but is becoming increasingly feasible, financially, for population-based association studies. Whole-exome sequencing only covers the exome, which is the collection of known exons[8] in a genome. Sequencing of just the exome is a cheaper method than whole-genome as exons represent 1-2% of the total sequence but are prime functional candidates.

1.2.3 Quality Control

GWAS data undergo dynamic quality control (QC) before analysis. The QC protocol is a critical step, conducted to remove any problematic samples and SNPs that can cause bias, increased false positive association rates and decreased power to detect associations. Like any experimental study involving patients and biological materials, there will inevitably be some missing genotype information and errors in genotype calling. QC can be separated into two parts: (i) SNP QC and (ii) sample QC. However, there is no standard order to follow in which each procedure should be carried out.

SNP QC

SNP QC procedures look to exclude low-quality SNPs based on the proportion of individuals for which a genotype has been called, deviation from HWE and MAF (Anderson et al. 2010). SNPs that have a low call rate, defined by a study-specific threshold are removed. Low call rates occur due to missing genotype calls in SNPs, which is

[8]Genes contain exons which are part of the protein-coding region in the DNA. Exons only comprise 1% of the genome and provide the most easily understood, functionally relevant information.

reflective of the problems with the array design or genotype calling methodology for that SNP. Assessing the distribution of the call rates by each marker, genome-wide shows outlying SNPs. A test for deviation from HWE is usually undertaken at each SNP, using Fisher's exact test at a study-specific significance threshold. Threshold changes are dependent on the number of SNPs under investigation. Extreme deviation from HWE is indicative of poor quality genotype calling at a SNP, for example, if heterozygous genotypes have been under-called compared to the common homozygous genotype. MAF is also used as a filter because the SNP quality tends to decrease with MAF. Low numbers in at least one genotype group also lead to lack of power to detect association amongst complex traits. The threshold used for filtering SNPs based on MAF is variable on the sample size of a study.

In order to conduct accurate imputation of samples, the allele calls from the study data and the reference panel data must be aligned to the same DNA strand (Verma et al. 2014). The differences between strands can arise due to the use of different genotyping platforms and calling algorithms for different sites of a multi-site study or between case and control groups. The SNPs that are not on the same strand must be "flipped" to the reference. Strand checks convert the reference allele to the alternative allele; however, all unresolved SNPs are usually discarded.

Sample QC

Sample quality control shares some similarities to SNP QC such as the exclusion based on low call rate, which is indicative of a poor quality DNA sample. Samples are removed due to the proportion of called genotypes that are heterozygous across autosomes (i.e. heterozygosity rate), which could represent sample contamination or inbreeding.

Discrepancies between the reported gender from external data and the genetic sex from the X chromosome genotype data can occur because of sample mix-up or errors in external data. The distribution of heterozygosity in males and females are different, as males will have no heterozygous genotype calls because they only have one copy of the

X chromosome. A routine initial screening to identify gender discordance is to plot each individual, separated by gender on an x-axis of the X-chromosome heterozygosity against the proportion of missing genotypes on the y-axis. Another method to identify gender discordance is using Wright's inbreeding coefficient, F, calculated from X-chromosome data. Software such as PLINK (Purcell et al. 2007) employs this check. For males, we expect the estimate to be close to $F = 1$, while for females, we expect close to $F = 0$. Allowing for low rates of genotyping errors, the threshold for males is $F < 0.8$ and $F > 0.2$ for females. Gender is sometimes adjusted for in association analyses; therefore it is essential that any misreported sex is excluded, along with any other collected phenotypic, clinical or external data for a sample that might be incorrect due to possible sample mix-up which will introduce bias and reduce power of the downstream association analyses.

1.2.4 Detecting and Accounting for Genetic Structure

Genetic structure arises due to relatedness between samples and population stratification. Population stratification is the presence of a systematic difference in allele frequencies between sub-populations in a study population, possibly due to different ancestry. Not accounting for structure can increase the false positive error rate as an association that is found could be due to the underlying structure of the population and not a trait associated locus. The simplest approach to assess false positives is using the distribution of observed test statistics summarised through a quantile-quantile (QQ) plot (see Figure 4.4). In the presence of population structure, the observed test statistics would be inflated over the expected under the null hypothesis of no association of the trait with SNPs genome-wide. Devlin & Roeder (1999) proposed correcting for population stratification using a method called genomic control, based on the observed association test statistics calculated under an additive model, (as described in Section 1.3.1) using an Armitage trend test. Deviation of observed test statistics from the null can be assessed by the genomic control inflation factor. A uniform correlation is then applied dividing observed test statistics by the inflation factor to adjust for population

structure.

Related or duplicate samples are a source of genetic structure. LD based pruning is usually undertaken to remove high LD regions or to avoid capturing too much variance of LD regions before checking for duplicated/related individuals. LD pruning is the process of creating a subset of markers that are in approximate linkage equilibrium with each other. For traditional association analysis (not family-based) samples should be independent of one another because the statistical methods used assume independent samples. Therefore related individuals defined as sharing the same alleles for a proportion of the genome are identified and discarded using the identity by state (IBS) matrix. Identical samples would have an IBS of 100%, with related samples sharing a high IBS. Another metric of relatedness that is often of interest is $\hat{\pi}$, known as identity by descent (IBD) estimation. This measurement is used to remove pairs of individuals that share a number of chromosomal regions identical by descent. Large values of $\hat{\pi}$ indicate relatedness between pairs of individuals. One sample from each of the related pairs is excluded based on a threshold of $\hat{\pi}$. The sample that is retained usually has the highest sample call rate.

The inclusion of samples that are ethnic outliers or the presence of population stratification are a source of confounding[9] due to genetic structure. The principle is that individuals who are geographically close together are likely to be more correlated in terms of genotypes than those who are far apart. Association analyses assume individuals are from a homogenous population background; otherwise, there can be an increase in type-I error rate.

Statistical techniques, such as principal components analysis (PCA), are essential in determining differences between populations and visually representing any genome-wide genotype differences between samples. PCA identifies ethnic outliers by merging the observed genotype data with reference genotypes from populations available from the international HapMap Project (Altshuler et al. 2010) or the 1000 Genomes Project

[9]Confounders are extraneous variables which confound the effect of the exposure on the outcome and which satisfy confounding criteria (Greenland et al. 1999).

(Auton et al. 2015). It is expected that the observed genotype data would cluster with the reference population data, with any outliers adrift from the clusters.

This technique can also be used to calculate principal components (PCs) that can be used to adjust for population stratification after the removal of ethnic outliers, by including them as covariates in the analyses of association (this can be adjusted for in the same way as the treatment covariate demonstrated in Section 1.3.1). The technique explained in detail by Price et al. (2006), is applied using eigenvalue decomposition of the genetic covariance matrix, deriving the PCs for the samples. An example of a PCA plot for the diversity across Europe is demonstrated by Novembre et al. (2008) (Figure 1.2). The most popular choice of software to undertake QC and the identification of genetic and population structure are both versions of the PLINK (Purcell et al. 2007, Chang et al. 2015) software.

1.2.5 Single-SNP Association Testing

In order to analyse the genome, SNPs are determined for all individuals. Each SNP passing QC is analysed separately for association with the trait of interest. Association analysis can be conducted using a number of statistical models usually within a regression framework assuming a particular mode of inheritance. These statistical models are outlined in further detail in Section 1.3. In most cases, the true genetic model is unknown and can be either additive, recessive or dominant. An "additive" genetic model is a very common assumption, whereby the effect of the heterozygous genotype is intermediate between the major and minor homozygous genotypes. Genotypes are coded according to the number of minor alleles carried, for example, assume the major and minor alleles at a SNP are A and T. The genotypes are therefore AA = 0, AT = 1 and TT = 2.

Imputation poses a different coding for genotypes. For each SNP a probability is derived for each possible genotype for an individual. The uncertainty in the genotype is modelled by averaging the three expected genotypes across the probabilities via a

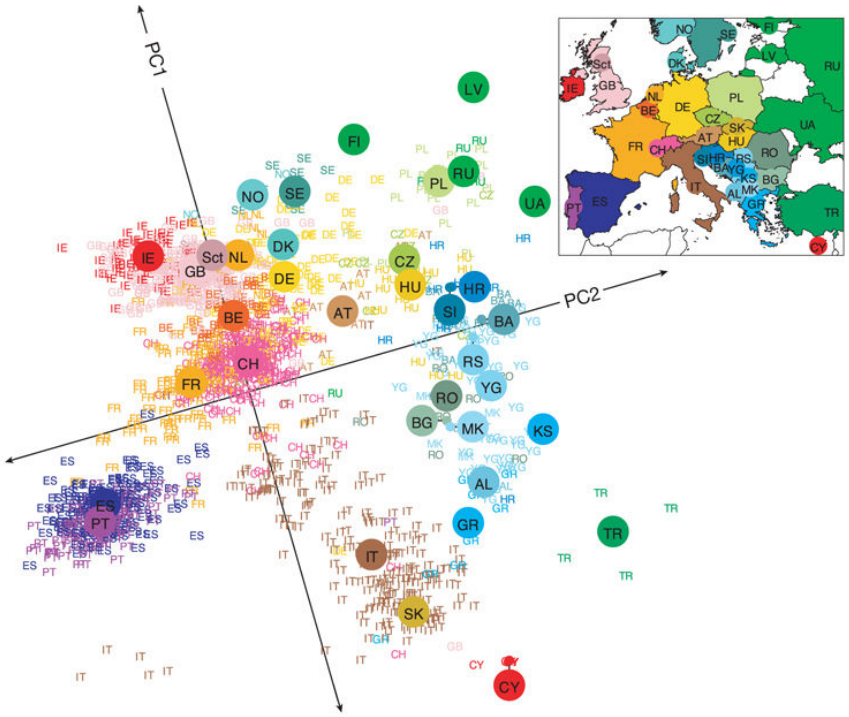


Figure 1.2: Population structure in Europe, defined by two principal components, from genome-wide SNP data in 1,387 individuals. Each small point (pairs of letters) correspond to an individual, plotted for the first two principal components, and coloured according to the country from which they were ascertained. The large circles correspond to medians for the first two principal components for individuals from each country. The projection has been rotated to emphasise the correlation with European geography (inset). Country abbreviations: AL, Albania; AT, Austria; BA, Bosnia-Herzegovina; BE, Belgium; BG, Bulgaria; CH, Switzerland; CY, Cyprus; CZ, Czech Republic; DE, Germany; DK, Denmark; ES, Spain; FI, Finland; FR, France; GB, United Kingdom; GR, Greece; HR, Croatia; HU, Hungary; IE, Ireland; IT, Italy; KS, Kosovo; LV, Latvia; MK, Macedonia; NO, Norway; NL, Netherlands; PL, Poland; PT, Portugal; RO, Romania; RS, Serbia and Montenegro; RU, Russia, Sct, Scotland; SE, Sweden; SI, Slovenia; SK, Slovakia; TR, Turkey; UA, Ukraine; YG, Yugoslavia. Originated from Novembre et al. (2008).

"genotype dosage" (see Eq. 1.2).

Within the statistical regression framework for association testing, confounding factors such as covariates and PCs can be adjusted for. These factors may contribute to the association between SNPs and the trait of interest.

1.2.6 Rare Variant Association Testing

NGS technologies have become available in recent years and can sequence a large number of samples, making much of the human variation accessible, including rare genetic variants, which are not typically captured by GWAS genotyping arrays (even after imputation). Two common limits used to distinguish rare genetic variants from common variants is a MAF less than 5% or 1%. The movement towards NGS data has allowed us to look more deeply into rare genetic variants and investigate their role in complex traits.

Pre-2015 many investigations had raised hopes that rare variant association studies (RVAS) would yield a large number of strong effect variants for the purpose of personalised medicine[10], consequently resulting in a plethora of new drug targets. However, the expectations have not been met, whereby most rare variants identified to date have modest effect sizes (Auer & Lettre 2015). However, with the increased coverage of the genome using various technological advances such as whole-genome sequencing, the likelihood of finding statistically relevant associations potentially can increase. To date, most findings have been from GWAS that target common variants, that rarely succeed in implicating specific genes (i.e. in non-coding regions) to common diseases, which limits their importance in applications such as drug development. However, rare variants could be the primary drivers of common diseases (Cirulli & Goldstein 2010).

It is not possible to analyse rare variants with traditional methods used for GWAS as they have insufficient power to detect associations with variants with MAF less than 5%. Currently, the best strategy for analysing rare variants is to combine them within units of association termed "gene-based" analyses, defined using gene annotations, genomic coordinates or functional characterisation (see Chapter 6).

[10]Personalised medicine is a treatment that is tailored to an individual based on characteristics such as demographics, clinical measurements and genetic factors. This area seeks to end the traditional 'one size fits all' approach to treatment, in the hope of developing better healthcare by maximising benefit and minimising harmful events.

1.2.7 Visualisation of Results

Association analysis results are best visualised through a Manhattan plot. These plots show SNPs represented by a point, plotted according to -log_{10} p-values (y-axis) against genomic location (x-axis). The benefit of this plot is that it helps distinguish chromosomes and SNPs in the presence of millions of statistical tests (see Figure 4.3). The tall towers of points clustered together represent SNPs in potentially strong LD with each other. An in-depth look at a particular location of the Manhattan plot, to investigate the pattern of association signals can be visualised using LocusZoom (Pruim et al. 2010). Each SNP is coloured according to the strength of LD with the lead[11] SNP, based on reference panel data, such as from the 1000 Genomes Project. Furthermore, local genes are defined below the plot (Figure B.1).

As mentioned in Section 1.2.4 even modest levels of confounding can distort the null distribution and overwhelm a small number of true associations. A QQ-plot is also mandatory to evaluate whether there is any evidence of genomic inflation, analytical approach bias or presence of population substructure.

1.2.8 Pharmacogenetics

The field of pharmacogenetics aims to identify genetic variants associated with drug efficacy and safety. Both pharmacogenetics research and GWAS, together will continue to help underlie genetic biomarkers and their relationship with drug response and metabolism[12] (see Table 2.1).

Individuals sometimes respond differently to treatments and alternative doses of a drug, which may reflect genetic differences between them. Therefore pharmacogenetics is key to understanding adverse drug reactions and efficacy to optimise treatments and improve patient care (Innocenti 2005). There is a rapidly expanding list of genetic

[11]The lead SNP (usually identified as the most significant SNP associated with the outcome) is a reference SNP whereby other SNPs (potentially causal) in the region can be identified through LD.

[12]This process is typically responsible for converting drugs to compounds that are easily absorbed and excreted.

variants that lead to altered drug responses. The GWAS Catalog (MacArthur et al. 2017) as of January 2018 shows a total of 225 SNP associations with a response to a drug.

Variants common and rare can now be assessed for their effects on response to treatment. Therefore, the number of genetic loci that predict response for efficacy and safety to specific drugs will continue to increase. This information will permit better-designed clinical studies, with more predictable outcomes (Kamb et al. 2013).

1.3 Fundamentals of Survival Analysis

The focus of most GWAS has primarily been on binary and quantitative phenotypes since these are the most commonly encountered outcomes when studying complex traits. However, more population-based cohort studies are being undertaken, which provide long follow-up in combination with genetic data. Therefore, it is now possible to not only analyse the risk of developing disease through a case-control outcome but also the time to particular disease onset or another event. Specifically, in pharmacogenetics, outcomes are often the time until the occurrence of disease remission or treatment withdrawal due to an adverse drug reaction. The identification of genetic variants associated with time to a key event (survival, death, relapse) after a clinical intervention has the potential to have a major impact on drug and dose choice by improving the benefit/risk ratio for a range of human diseases with substantial population health burden.

The most powerful analytical approaches for testing association between genetic variants and these outcomes are to model the time to the occurrence of the event, adopting survival analysis methodologies. Methods implementing a Cox proportional hazards (PH) model are already extensively applied within candidate gene studies or small-scale genetic association studies.

It is important not only to find associations between genetic variants and the outcome of interest but also to quantify the impact of the effect on a trait. Exploring the effects

on survival of a group of patients depends on the values of many explanatory variables, which are recorded for each patient throughout a study. In the analysis of survival data, interest lies on the risk or hazard of an event at any time after the study begins (Collett 2003). Two important functions which determine the distribution of the event times are the survivor and hazard functions. The survivor function, denoted by $S(t)$, is the probability of survival to time t, whereas the hazard function, $h(t)$, is the probability that at any given moment, the event will occur, given that it has not already done so.

Time-to-event (TTE) studies have an added complexity when modelling the event of interest because it has usually not yet occurred for all individuals at the end of follow-up or has occurred at an unknown time before follow-up. These individuals with incomplete survival times are called censored observations. Many different types of censoring can occur. However, the most common is right censoring, which occurs when a study ends before all individuals have experienced the event of interest. Censoring can occur due to a number of reasons such as a patient dropping-out from a study. Together with censoring, another unique feature of TTE studies is that the observed event times for a study population are often highly skewed, following a wide range of non-standard statistical distributions.

1.3.1 Statistical Models

There are many different survival models available for the analysis of TTE data, including; non-parametric[13], semi-parametric and parametric models. The purpose of this section will be to define the statistical methodology used throughout the thesis, covering semi-parametric, parametric and alternative dichotomous approaches. Any proposed extensions to these methods have been defined in the subsequent chapters.

[13]Non-parametric tests such as the log-rank test compare the survival of two groups, providing a p-value of significance, however it does not quantify the difference between the two groups.

Cox Proportional Hazards Model

The concept of PH is defined as the hazard of an event at a particular time point for individuals in one group being proportional to the hazard at the same time point for another group.

$$h_1(t) = \phi h_2(t) \qquad \text{(Eq. 1.1)}$$

ϕ is a constant known as the hazard ratio (HR) of the event occurring at any time for an individual. For example, in Eq. 1.1 we are assuming that the hazard for an individual with a heterozygous genotype, $h_1(t)$, is proportional to the hazard for an individual with a homozygous genotype, $h_2(t)$, constantly over the study period.

The Cox PH model is the most widely used approach when modelling TTE outcomes. It is a semi-parametric model where the HR takes a parametric form regarding the regression coefficients, but the baseline hazard is unspecified (Cox 1975). The model is defined using a partial likelihood function, rendering it computationally beneficial with an additional advantage of being able to adjust for covariates. A disadvantage of this model is that the distribution of survival times is unknown. In cases where the PH assumption is not valid, other analysis models or extensions to the Cox PH model should be considered. Furthermore, the Cox PH model is known to have poor properties when the sample size is small, or when the risk factor is imbalanced, i.e. the sample size is small in one risk group (Wang et al. 2010).

Consider a study investigating the association of a TTE outcome with genetic variants in a sample of unrelated individuals. Let t_i, denote the TTE for the i'th individual and their additional covariates by the vector $\hat{\mathbf{x}}_\mathbf{i}$. Also, let G_{ij}, denote their genotype at the j'th SNP of interest. Under an additive genetic model, the genotype of the i'th individual is coded as $[0, 1, 2]$, defined by the number of minor alleles they carry at the variant. Imputed genotypes are modelled as an additive effect using the probabilities p_{ij1} (heterozygote) and p_{ij2} (minor homozygote):

$$G_{ij} = p_{ij1} + 2p_{ij2} \qquad \text{(Eq. 1.2)}$$

Under the assumption of PH, we can express the hazard of the event occurring at some time t for the i'th individual given their genotype at a j'th SNP, conditional on them not yet having experienced the event by:

$$h_i(t) = h_0(t)e^{\beta_s G_{ij} + \hat{\beta}_{\mathbf{x}}\hat{\mathbf{x}}_\mathbf{i}} \quad \text{(Eq. 1.3)}$$

In this model (Eq. 1.3), $h_0(t)$ is the baseline hazard at time t, and the parameters β_s and $\hat{\beta}_{\mathbf{x}}$ correspond to the log-relative hazard for each copy of the minor allele at the SNP, and a vector of covariate effect(s), respectively. The partial likelihood is given by:

$$\mathcal{L}(t|\beta_s, \hat{\beta}_{\mathbf{x}}, G, \hat{\mathbf{x}}) = \prod_i^n \frac{e^{\beta_s G_{ij} + \hat{\beta}_{\mathbf{x}}\hat{\mathbf{x}}_\mathbf{i}} c_i}{\sum_{j \in R(t_i)} e^{\beta_s G_{ij} + \hat{\beta}_{\mathbf{x}}\hat{\mathbf{x}}_\mathbf{i}}} \quad \text{(Eq. 1.4)}$$

where, c_i is an indicator taking the values 1 if the event occurred at time t_i, and 0 if the observation was censored. In this expression, $R(t_i)$ denotes the risk set at time t_i, corresponding to individuals who have not yet either experienced the event or been censored.

The interpretation of the parameter estimates from the Cox PH model is the log relative hazard associated with a one-unit increase in the covariate, which means that when we see a positive estimate using the Cox PH model it is an increased hazard of the event.

Weibull Regression Model

The Weibull regression model is a parametric survival model that makes an assumption about the statistical distribution of the data and has completely specified hazard and survivor functions. The model is beneficial when the HR is not proportional over time or the data have an accelerated failure time (AFT) feature, whereby the effect of the covariate is multiplicative on the time scale and it is said to "accelerate" survival time. Under these conditions we might expect the power of these approaches to be greater than the Cox PH model. However, a potential drawback to this is that you need to make a correct assumption on the underlying distribution and shape of the hazard function, otherwise the results may be misleading.

This model is most commonly used in situations where it is of interest to estimate the mean of the survival distribution, the survivor function and acceleration factor (AF). The AF evaluates the effect of predictor variables on the survival time. The effect of variables in the Weibull AFT model is to accelerate or decelerate time by some factor. An additional benefit specific to the Weibull model is that an estimate for both the HR and AF can be calculated.

The Weibull regression model for right censored observations can be derived using the Weibull distribution density and survivor functions. Let t_i, denote the TTE for the i'th individual and their vector of covariates by $\hat{\mathbf{x}}_\mathbf{i}$. Also let G_{ij}, denote their genotype at the j'th SNP of interest coded under an additive dosage model for the minor allele. Within this framework, $f(t) = b_i a t^{a-1} e^{-b_i t^a}$ is the density function and $S(t) = e^{-b_i t^a}$ is the survivor function, where a is the shape parameter and b_i is the scale parameter. We parametrise the scale and shape parameters to incorporate the unknown regression coefficients.

$$b_i = \frac{1}{e^{\beta_0 + \beta_s G_{ij} + \hat{\beta}_\mathbf{x} \hat{\mathbf{x}}_\mathbf{i}}} \tag{Eq. 1.5}$$

$$a = \frac{1}{\sigma} \tag{Eq. 1.6}$$

The parameters β_0, β_s and $\hat{\beta}_\mathbf{x}$ correspond to the log-relative change in time for the intercept, each copy of the minor allele at the SNP, and a vector of covariate effect(s), respectively. The likelihood of the observed time to event data under the Weibull model is then given by:

$$\mathcal{L}(t|\theta) = \prod_i^n \left\{ [f(t|\theta)]^{c_i} \, [S(t|\theta)]^{1-c_i} \right\} \tag{Eq. 1.7}$$

$$c_i = \begin{cases} 1 & \text{individual has had event} \\ 0 & \text{individual is censored} \end{cases}$$

To obtain maximum likelihood estimates for the set of model parameters $\theta = \{\beta_0, \beta_s, \hat{\beta}_\mathbf{x}, \sigma\}$, we maximise Eq. 1.7 and use the Newton-Raphson iterative method. To obtain reasonable updates of the model parameters the shape paramter σ is updated

very gradually in increments of 0.1.

$$\theta_{N+1} = \theta_N - H^{-1}\Delta$$

Where, H is the Hessian matrix, Δ is the first order derivative vector, and θ_N are parameter estimates in the N'th iteration. From the maximum-likelihood parameter estimates, we calculate a z-statistic for each regression co-efficient, given by:

$$z = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

$$p = 2\left(1 - \int_{-\infty}^{|z|} \frac{2}{\sqrt{\pi}} e^{\frac{y^2}{2}} dy\right) \quad \text{(Eq. 1.8)}$$

The estimates from the Weibull regression model are interpreted as the log-relative change in TTE (or AF) associated with a one-unit increase in the covariate. This means that when we see a positive estimate using the Cox PH model (increased hazard of event) we expect a negative estimate using the Weibull regression model (earlier occurrence of event). The p-value for each regression parameter is calculated using Eq. 1.8. Alternative tests to calculating the significance of explanatory variables in a model or comparing two models exist, such as the Wald test and likelihood ratio test (LRT). The Wald test statistic is given by:

$$W = I(\theta)\left[\theta - \theta_{Null}\right]^2 \quad \text{(Eq. 1.9)}$$

$I(\theta)$ is the expected Fisher information matrix. θ_{Null} is the proposed values (null model). The assumption is that the difference between θ and θ_{Null} will be approximately normally distributed. The Wald test statistic can be used to calculate the p-value for each model parameter in the Weibull, Cox PH or another regression model. The single parameter test statistic is compared to a chi-squared distribution:

$$W^2 = \frac{(\hat{\beta} - \beta_{Null})^2}{Var(\hat{\beta})} \sim \chi_1^2 \quad \text{(Eq. 1.10)}$$

Where, β_{Null} is usually 0. The LRT is used to compare two statistical models (i.e. the null and alternative), given by:

$$2(\ell(\theta) - \ell(\theta_{Null})) \quad \text{(Eq. 1.11)}$$

The p-value is derived using the difference between model log-likelihoods (Eq. 1.11). The probability distribution of the test statistic is approximately a χ^2 distribution with degrees of freedom equal to the number of free parameters between the alternative and null models.

Logistic Regression Model

Alternatively, to modelling the event times, we can dichotomise the outcome of individuals at the end of the study. Let y_i denote the dichotomised outcome of the i'th individual, given by $y_i = 1$ if the event has occurred, and by $y_i = 0$ if not. Individuals who are censored before the end of the study are excluded from this analysis since it is unknown whether the event has occurred or not and thus are treated as missing. Within a logistic regression framework, we can model the log-odds of the occurrence of the event by:

$$logit(y_i) = \beta_0 + \beta_s G_{ij} + \hat{\beta}_{\mathbf{x}}\hat{\mathbf{x}}_{\mathbf{i}} \qquad \text{(Eq. 1.12)}$$

In this expression, β_s corresponds to the log-odds ratio of the minor allele relative to the major allele at the SNP, and $\hat{\beta}_{\mathbf{x}}$ represents a vector of covariate effect(s). We can form a likelihood ratio test of association of the SNP, j, with the outcome by maximising the unrestricted model Eq. 1.12 and comparing with that under the null model, for which the allelic log-odds ratio, β_s, is zero.

1.4 The Importance of Statistical Genetics Software

In the era of large-scale GWAS of thousands of individuals at millions of SNPs, datasets can eclipse the size of hundreds of terabytes. Computational tools are key for handling and processing data efficiently. These tools are mainly developed for Linux operating systems (O/S) as this is largely the O/S run on high-performance computing (HPC) servers. There are a variety of programming languages available that are used in the production of statistical genetics software. Command line genetic data analysis software such as PLINK (Purcell et al. 2007) and SNPTEST (https://mathgen.stats.ox.

ac.uk/genetics_software/snptest/snptest.html) are written in C++ with a large number of new software pipelines such as HAIL (Seed et al. 2017) that are developed using Python and Scala. These languages are rising in popularity because of the syntax readability, their ability to handle scientific datasets and provide a framework for machine learning.

Software to perform power calculations is usually made to cater for multiple O/S, utilising a graphical user interface (GUI). The reason for this is that they perform less computationally intensive tasks and should be made with the specifications for those without knowledge of the command line to use with ease. Power calculators seek to determine the sample size and statistical power for a particular study depending on a number of specified parameters. Many power calculators are web-based, such as the genetic power calculator (GPC) (Purcell et al. 2003). Software like GPC allows all users with access to the internet the ability to run the program without the need to download and install the program on a local computer.

With the introduction of NGS, genetic datasets have undergone many changes. Alongside the change in sequencing and genotyping platforms, the data volume increases, and different file types that store the data are produced. Due to this, computational tools need to evolve with these rapid changes, such as the movement from the genotype (GEN)[.gen] file format to the variant call format (VCF)[.vcf] for sequence-based genotypes. Software tools are changing the landscape of genetics research. They are making all stages of the genetic variant discovery process more convenient to undertake, from study design through to interpretation of analysis results.

1.5 Thesis Objective and Structure

The core aim of this thesis is to develop and evaluate statistical methodology for GWAS with TTE outcomes, specifically but not limited to the field of pharmacogenetics. The methodology is implemented within a set of computational tools consisting of a data simulator and power calculator, and both single- and rare-variant analysis programs. Secondary aims are to conduct simulation studies for all developed methodology and

software with a final application to the Pharmacogenetics of Acute Coronary Syndrome (PhACS) study, which aims to identify genetic risk factors for the recurrence of cardiovascular events after treatment. The motivation behind this research is to draw attention to the underdevelopment of methodology and software for GWAS of common and rare variants with TTE outcomes. Correct modelling of these outcomes within GWAS has important implications in determining associations between genetic biomarkers and traits or diseases. The chapters in this thesis can be separated into four main topics of interest: literature review, methodology conception, software development and application to data.

Chapter 2 provides a review of key elements of pharmacogenetic study designs and methodology. An evaluation is conducted using simulation in R, of the power to detect an association between SNPs and TTE outcomes across a range of pharmacogenetic study designs whilst comparing alternative regression approaches. Comparison of statistical power is made using: (i) a Cox PH model; and (ii) a logistic regression framework with a dichotomised outcome at the end of the study. The investigation incorporates detailed simulations and empirical evaluation in a candidate gene study of anti-epileptic drug response with SNPs mapping to/near the *ABCB1* gene (Leschziner et al. 2006). The purpose of this is to identify scenarios for which the difference in power between the analytical models is minimised.

Chapter 3 briefly reviews current software for simulating genetic data and performing power calculations. From this, a detailed description is made of the software SurvivalGWAS_Power which performs power calculation for pharmacogenetic TTE studies over a range of designs and analytical models. Use of the software is demonstrated through a basic comparison of Cox PH and Weibull regression models.

Chapter 4 provides a comprehensive review of GWAS with TTE outcomes, methodology development and current software for a variety of different outcomes. A detailed outline of the software SurvivalGWAS_SV is given, describing the implementation of methodology and processing algorithms, through to examples of use. The example is

an application using simulated SNP data, undertaken for the purpose of testing software efficiency and the appropriateness of the statistical methods.

A novel methodological framework for gene-based tests of association, with implementation within the program rareSurvival, is introduced in Chapter 5. The methodology conception is explained along with the program's algorithmic pipeline. The program is tested using a simulation study based on exome array genotype data from two Swedish cohorts.

The primary application of the novel methodology and software is to the PhACS study in Chapter 6. The PhACS study is a prospective pharmacogenetic cohort of 1470 patients who had a cardiovascular event followed-up 48 months after hospital discharge. The data is analysed with both SurvivalGWAS_SV and rareSurvival, with the main objective of identifying SNPs and genes associated with time to recurrent cardiovascular event and all-cause mortality. Testing the software under realistic conditions, provided insight into the efficiency of both computational analysis tools.

Concluding remarks and further research proposals are outlined in Chapter 7. The proposals cover improvements for each of the three computational tools, development of novel statistical methods for more complex outcomes and study designs. Notably, the impact of the research contained within this thesis and the future outlook and direction of the field is discussed.

CHAPTER 2

EVALUATION OF METHODOLOGY FOR PHARMACOGENETIC "TIME-TO-EVENT" STUDIES

2.1 Overview

Methodology for modelling time-to-event (TTE) data has been extensively developed over the years within medical research, from non-parametric procedures such as the log-rank test to accelerated failure time models (Section 1.3). In the context of single variant analysis of genomic data where genetic variants are considered as predictors in the model to estimate hazard ratios, the same methodology can be applied and has been done so over many years within pharmacogenetic candidate gene studies (Table 2.1). However, this has not been the case for genome-wide association studies (GWAS), due to the number of single nucleotide polymorphisms (SNPs), samples and the genotype uncertainty from imputation to consider. In this instance, the current methodology is not the limiting factor, rather the availability of analysis software implemented with sophisticated data handling algorithms to perform tests of association as quickly and efficiently as possible. This topic is further discussed in Chapter 3 of this thesis.

The traditional approach to the analysis of TTE data is through survival modelling. Some of these models explicitly account for the most distinctive feature of nearly all TTE studies: censoring. There are many different types of censoring, including right, left, administrative or interval. These can occur collectively or individually for many reasons, for instance, the study ending before the patient experiences the event of interest or patient drop-out before the end of the study. The latter is a widespread occurrence in TTE studies, because the trials often extend over a long period, waiting for individuals to have the event of interest. In pharmacogenetic studies, the outcome of interest is usually TTE after treatment intervention, where the event could be death, disease remission, or the occurrence of an adverse drug reaction.

The Cox proportional hazards (PH) model is the most popular choice for analysis when examining TTE data. This notion is documented in many pharmacogenetic studies with survival phenotypes (Table 2.1). However, there are instances where a study records TTE outcomes, and follow-up data are collected, but a decision has been made to look at a dichotomous outcome while removing individuals from analysis who would have otherwise been censored in a TTE analysis.

2.1.1 Dichotomising Time-to-Event Outcomes

In the past, dichotomisation of the TTE outcome was largely undertaken for genome-wide analyses, since available software packages are mostly limited to binary and quantitative traits. One approach to circumvent the problem of lack of TTE software availability is to consider the occurrence of the event as a dichotomous outcome at some fixed time point, such as the end of the study. Individuals in which the event has occurred are considered as "cases", while those in which the event has not occurred are considered as "controls" (Ji et al. 2013, Speed et al. 2014). Nevertheless, this approach would be expected to result in a loss of power to detect association with SNPs compared with direct modelling of the TTE outcome because: (i) the event times are not directly considered, thereby losing information; and (ii) the binary outcome cannot allow for censoring before the end of the study, in which case individuals will be treated as "missing" observations. Even though these individuals are classified as missing data because the TTE has not been observed, they are valuable as the observation that they went event free over a period is itself highly informative.

George et al. (2014) provide a few introductory examples on why we use survival analysis. An example will be if a study has found that the final observed proportion of events between treatment groups is identical. However, if one group had all events occur shortly after randomisation, while the other had no events until just before the end of follow-up, then the two treatments would logically be considered to have different clinical effects despite the same proportions of events at the end of follow-up.

A GWAS conducted by Ji et al. (2013) looked at depression disorder, and the treatment outcomes were collected as a score over an eight-week outpatient clinical trial. They use the logistic regression model to analyse the outcomes response and remission as binary. The analysis did not yield any SNPs reaching genome-wide significance $\left(p < 5 \times 10^{-8}\right)$. Nonetheless, they found 14 promising SNPs with one associated with treatment response. This study had the information available for a TTE analysis to be conducted with outcomes such as time to response and remission, but these data were not used in a survival analysis. The article also states that two strategies were used to analyse the primary outcomes: (i) "the primary analyses included only individuals that were evaluated at the 8-week visit"; and (ii) "the secondary analyses were performed with outcomes based on the final visit [...] these analyses included subjects who had completed the full 8-week study, as well as those who dropped out of the study before the 8-week assessment". This sentence means that those who dropped out of the study would be classified as censored observations. However, the logistic regression model used would not account for the observations in the primary analysis, and in the secondary analysis, those that dropped out would be non-events. Examples of this approach can also be found in candidate gene studies such as Charland et al. (2014), Clarke et al. (2014) and Lohoff et al. (2013).

There is extensive literature describing the relative lack of power of binary analyses of TTE outcomes, but comparisons have not been made in the context of pharmacogenetic studies. Although there might be power losses by simplifying the outcome, there may be substantial savings in computational runtime and resources.

2.1.2 Objectives

This chapter seeks to address the following questions:

1. What are the most commonly used pharmacogenetic study designs?

2. What methodology and software are used for analysis?

3. Under which pharmacogenetic settings is the loss in power smallest from assum-

ing a dichotomised outcome instead of applying a survival analysis approach?

In this chapter, published pharmacogenetic studies are reviewed to summarise the most widely used designs and analytical approaches. Simulations have been undertaken to investigate the circumstances under which dichotomisation of a TTE outcome, has minimal loss in power compared to traditional survival models.

2.2 More Than Five Years of Pharmacogenetic Studies

The field of pharmacogenetics is becoming increasingly widespread year after year while moving through the GWAS era. Studies are being published for both candidate gene studies and GWAS. Together, they complement one another in the pursuit of a common goal: personalised medicine.

To understand the different types of study designs and analysis methodology used within pharmacogenetics, an evaluation of a total of 42 studies was carried out. This began by conducting a search on PubMed (https://www.ncbi.nlm.nih.gov/pmc/) and Google (http://www.google.co.uk) using a combination of the following keywords; "genome-wide", "survival", "pharmacogenetics", "pharmacogenomics" and "time to". Further searches were conducted using the keywords directly within The Pharmacogenomics Journal (https://www.ncbi.nlm.nih.gov/labs/journals/pharmacogenomics-j/) and Pharmacogenomics (https://www.futuremedicine.com/journal/pgs) website search bars.

After the literature search, the exclusion criteria involved first sorting the papers by relevance and then filtering the papers based on publication date, to include only published articles from the last five years[1]. Figure 2.1 provides a more detailed representation of the number of articles included at every stage of the review process. The reason for using the latest articles was to eliminate designs that may not be in use anymore, and by understanding the current study designs, helps us gather ideas about future designs. All articles included were between 1 January 2012 to 13 November 2015. Articles

[1]Initial search date: November 2014; Secondary search date: June 2015; Final search date: November 2015.

beyond this point are discussed in subsequent chapters. If the search indicated that the manuscript completely focused on a pharmacogenetic GWAS application with survival or TTE phenotypes, then these would match exact criteria for inclusion. However, there are very few of these papers. Therefore papers which focussed on candidate gene pharmacogenetics with TTE phenotypes were also reviewed. The remaining papers which covered only pharmacogenetics within a candidate gene or GWAS with a binary outcome were examined based on their study design section. For example, we included papers that highlighted an interesting setting that was not present in the other papers, but which could still be used in TTE studies.

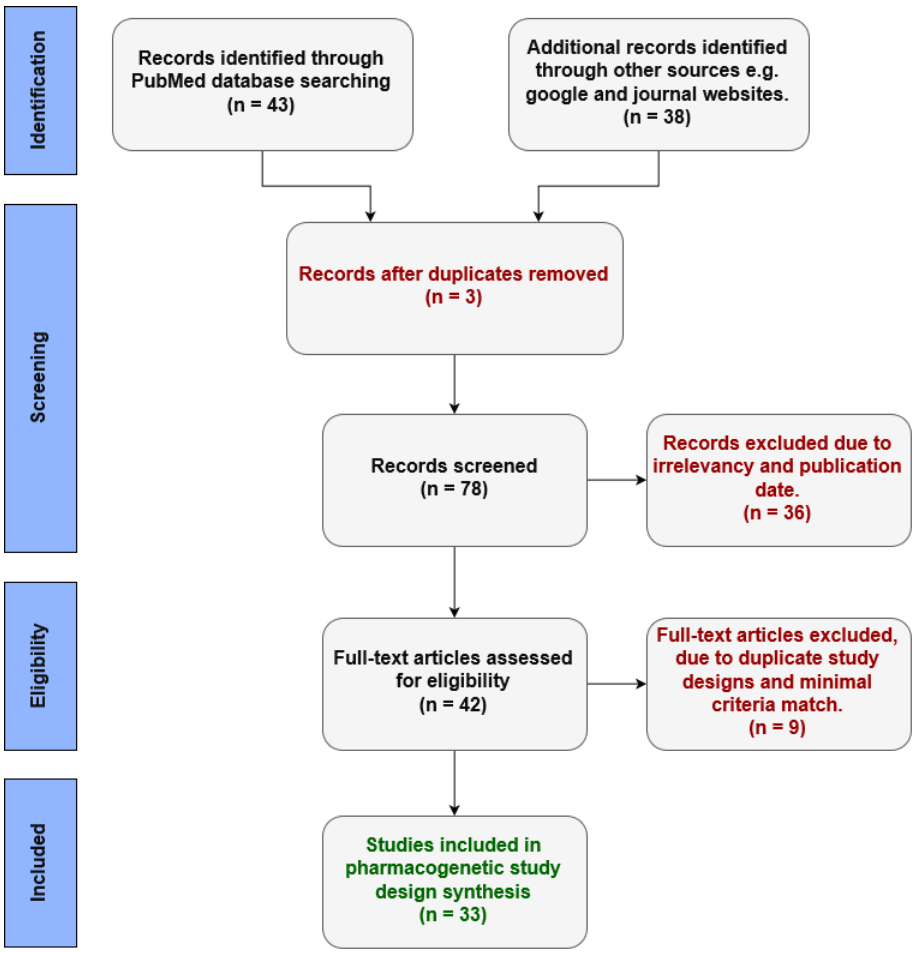


Figure 2.1: Literature review eligibility flowchart for pharmacogenetic studies. Blue boxes indicate process stages. Red coloured text represent excluded articles, and the green text indicates the final article inclusion.

Table 2.1 presents a summary of the most notable articles examined in the literature review. It highlights the key features of each study design and statistical analysis

protocol. In total 33 studies were chosen for further examination of which 25 used a candidate gene approach, 8 were a GWAS of common variants, and 20 studies included a TTE outcome. The following classification criteria were used to assess a study's relevance for inclusion:

1. GWAS or candidate gene study?
2. Common or rare variants?
3. Was the data imputed?
4. Number of SNPs.
5. Disease of interest.
6. Treatment/Intervention.
7. Study design.
8. Outcome.
9. Analysis model.
10. Software used.

From these articles, a few essential study design features at the foundation of most pharmacogenetic studies can be identified. Key characteristics of TTE and pharmacogenetic studies include scenarios with: (i) a recruitment period, defined as the length of time within a study where individuals are recruited, and phenotype information is collected; (ii) follow-up time; (iii) treatment intervention; (iv) multiple drug doses; and (v) right censoring.

The length of a study from patient recruitment to the end of follow-up can determine the number of events and censored observations that occur. For example, if the follow-up time for each individual in a study is after a short period of time for diseases such as cancer, where interest lies with progression-free survival, then all patients are alive without disease progression. This will result in no events across genotype groups. In such cases, the TTE contains much more clinical information than whether or not the event occurred.

Study	**Type**	**Variants**	**Disease**	**Study Design**	**Phenotype**	**Method & Software**
Innocenti et al. (2012)	GWAS	484,523 SNPs after QC.	Pancreatic cancer	351 patients treated with test treatment or placebo. Given on selected days of a 28-day cycle.	OS.Treatment was stopped for progressive disease, adverse events, or patient withdrawal.	Log-linear two way multiplicative CPHM. Additive genetic model. R '*survival*' package.
Ray et al. (2015)	Candidate gene	6 (*ABCB1*)	Chronic depression	83 patients given 8-12 week pharmacotherapy. Anti-depressant treatment. Multiple drugs and dosages.	Time to remission. Individuals who crossed over to another drug before remission were censored.	CPHM. Software unknown.
Ashare et al. (2013)	Candidate gene	*APOE* $\epsilon 4$	Smoking cessation	917 patients in placebo controlled trial. Adaptive treatment regimen with increases in drug dose.	Time to 7 day failure. Genotype x age, and genotype x treatment arm interactions.	CPHM. STATA 12.0.

Table 2.1 continued from previous page

Study	**Type**	**Variants**	**Disease**	**Study Design**	**Phenotype**	**Method & Software**
Koutras et al. (2014)	Candidate gene	*VEGF*	Breast cancer	Multiple treatments and doses. Treatments adapted during 12 weeks of follow-up depending on severity of disease.	OS. PFS.	Multivariate CPHM. SAS 9.3.
Depta et al. (2015)	Candidate gene	*CYP2C19* genotype	Myocardial infarction	12 month follow-up. Patients treated with clopidogrel.	All cause mortality. Time until cardiac rehospitalisation.	CPHM. SAS 9.2.
Absenger et al. (2014)	Candidate gene	*CCND1*	Colon Cancer	264 patients. 2 groups of different stage cancer patients treated with chemotherapy.	TTR. Censored at the time of death or at the last follow-up if the patient remained tumor recurrence-free.	CPHM. SAS 9.2.
Ji et al.. (2013)	GWAS	532 877 SNPs	Major depressive disorder	8 week outpatient clinical trial.	Treatment response. 8 week remission and response.	Logistic regression model. R & PLINK.

Table 2.1 continued from previous page

Study	Type	Variants	Disease	Study Design	Phenotype	Method & Software
Han et al. (2014)	GWAS	334,127 SNPs	Lung cancer	All patients received 2 treatments, different doses on various days.	OS. PFS. Patients lost to follow-up, were censored on the date of last contact.	Multivariate CPHM. PLINK 1.07 & SAS 9.1.3.
Uppugunduri et al. (2014)	Candidate gene	4 *CYP* genotypes.	Stem cell transplant	Age dependent treatment dose. Dosage adjusted at 5'th administration.	EFS was defined as the time from the day of transplant to the occurrence of any event of interest.	Multivariate CPHM. SPSS 19.0.
Fernandez-Rozadilla *et al.* (2013)	GWAS	497,366 SNPs	Colerectal cancer	Colon cancer patients receive 1 of 2 treatments whereas rectal cancer patients 1 of 2 treatments or a combination of both with radiation therapy.	Toxicity response. Patients receiving other treatment schedules were excluded from the study.	Logistic regression. PLINK.

Table 2.1 continued from previous page

Study	Type	Variants	Disease	Study Design	Phenotype	Method & Software
Pander et al. (2015)	GWAS	647,550 SNPs after QC.	Colerectal cancer	755 patients received multiple treatments, doses (adaptable) and delivery systems on different days.	PFS. SNP x treatment interaction.Treatment stopped at disease progression, toxicity or death.	CPHM. R '*survival*' package.
Sato et al. (2011)	GWAS	109,365 SNPs	Lung cancer	105 patients received a total of 308 cycles of treatment. Patients followed up until death or up to 5 years after treatment. All patients were followed up for more than 2.5 years.	OS was calculated from the date of recruitment to the date of death or the last follow-up.	CPHM. SAS 9.13.

Table 2.1: Summary of pharmacogenetic studies of special interest from literature review. This table is in no way a full description of the studies, but a general summary. Abbreviations: QC, quality control; OS, overall survival; PFS, progression free survival; TTR, time to tumour recurrence; EFS, event free survival; CPHM, Cox PH model; SNPs, single nucleotide polymorphisms; GWAS, genome-wide association study.

Cancer studies within pharmacogenetics are of interest because they investigate the effects of multiple treatment options and doses, and Han et al. (2014) is an excellent example of this. The study was a GWAS of survival in small-cell lung cancer patients treated with two chemotherapy options. The primary outcome was overall survival for 139 patients. After quality control, 334,127 SNPs were retained for analysis. The analysis was run using a multivariate Cox PH model in SAS. Even though this is a relatively small number of SNPs and sample size, a statistical analysis run in SAS can have many limitations. First, imputed genotypes increase the complexity of the data, thereby needing software to be able to read in the different file types from programs such as IMPUTE2 (Marchini et al. 2007). For imputed SNPs, the genotype is no longer an integer (0, 1 or 2), but is now equal to the expected allele count, previously referred to as dosage (Eq. 1.2). Second, as the number of variants and sample size increases, software such as SAS are not easily amenable to high-performance computing (HPC) clusters. Software such as SAS and R have a fast-growing implementation of packages covering a wide range of different data types, and as shown in Table 2.1, they are the most popular choices to run the Cox PH model. For the analysis of GWAS with binary and quantitative traits, SAS and R are rarely used especially since the introduction of bespoke software such as SNPTEST (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) and PLINK (Purcell et al. 2007, Chang et al. 2015). These tools offer efficient analysis and a "user-friendly" interface for those not familiar with coding environments. Table 2.1 shows three papers that have used PLINK for analysis. As explained earlier in Section 1.4, with these and other software available for binary outcomes, the truth remains that, if you simplify the survival outcome it can be analysed using a logistic regression model in these software packages with the benefit of computational runtime but at the cost of reduced power.

2.3 Simulation Study

Simulations were undertaken to compare alternative analytical approaches over a range of study designs, collated from the literature review in Section 2.2. The main aim of this

simulation study was to simulate a variety of realistic pharmacogenetic study designs while evaluating the power to detect an association between SNPs and TTE outcomes using alternative regression approaches. The study considers designs with censoring before the end of the study, treatment effects, SNP-treatment interaction effects, and a variable recruitment period.

This section compares the power for the analysis of event times: (i) under a Cox PH model; and (ii) within a logistic regression framework with a dichotomised outcome at the end of the study. Although initially, it is expected that the Cox PH modelling would be uniformly most powerful, the goal is to identify scenarios for which the difference in power is minimised and existing software, such as PLINK, could be utilised at minimum cost.

2.3.1 Procedure

Four commonly used design scenarios of pharmacogenetic GWAS were considered to evaluate SNP association with TTE outcomes. The scenarios are described in detail below in Section 2.3.2, and allowed for a variable end of study time and recruitment period, censoring before the end of the study, multiple treatment effects, and SNP-treatment interaction. All simulation scripts and analyses were written and performed in R 3.2.0 (R Core Team 2013). Data were simulated using statistical distribution functions, such as 'rweibull' and 'rbinom', for the different parameters discussed in Section 2.3.2. The '*survival*' package (Therneau 2015) in R was used to run the Cox PH model ('coxph' function) and the '*stats*' package (R Core Team 2013) was used for the logistic regression model ('glm' function). The output from all simulations are displayed as power plots, effect size and $-log_{10}\ p$-value plots.

2.3.2 Scenarios and Datasets

In all scenarios, a study undertaken for a maximum of 60 days was considered. Furthermore, the impact on the analysis by fixing the end of the study, Z, at 20, 30, 40, 50

or 60 days was examined. Patients who do not experience the event before the end of the study were assumed to be censored at that point. However, in some scenarios, it is imperative to allow for the possibility of censoring before the end of the study due to the occurrence of an adverse treatment reaction or other reasons for drop-out. In these

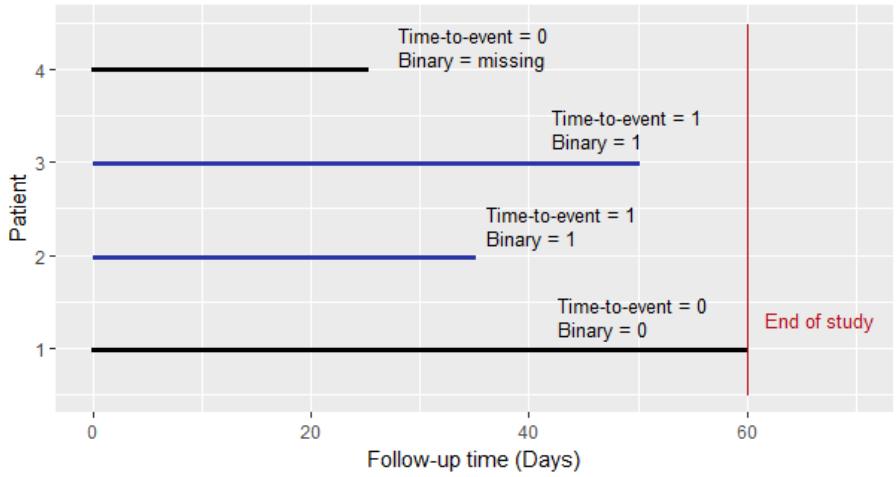


Figure 2.2: An example of right censoring for four patients. The blue lines and outcome (binary or TTE) equal to 1, represents the event has occurred. The black lines and outcome (binary or TTE) equal to 0, or missing, represents that the patient is censored. The vertical red line at 60 days signifies the end of the study.

scenarios, patients who are censored before the end of the study were excluded from the logistic regression analysis at the end of the study (Patient 4 in Figure 2.2) because it cannot be determined whether the event has occurred or not. Figure 2.2 depicts an example of how the TTE and binary outcomes are determined for four individuals.

For each scenario, consider a simulated SNP effect, ϕ_s, on the log-hazard of the occurrence of the event in the range of 0 (null model to evaluate false positive error rates) to 0.4. For each simulation, 1,000 replicates of data for a sample of 1,000 patients was generated. For each replicate, genotype data was simulated for the i'th patient, G_i, from a multinomial distribution with minor allele frequency (MAF) of 0.4, under the assumption of Hardy-Weinberg equilibrium and coded under an additive dosage model for the minor allele, $G_i = (0, 1, 2)$. Assuming PH, the time to the occurrence of the event, T_i, was simulated from a Weibull distribution with shape parameter 1 and scale parameter b_i. b_i is dependent on the genotype G_i, SNP effect ϕ_s, vector of log-hazard

treatment effects $\hat{\phi}_{\mathbf{x}}$ for the vector of treatments $\hat{\mathbf{x}}_{\mathbf{i}}$ and other scenario-specific factors outlined below. The shape parameter of the Weibull distribution controls whether the rate of events is increasing, decreasing or constant over time, while the scale parameter determines the dispersion of the distributional values.

$$T_i = Weibull\left(1, b_i = d_0(t)e^{\phi_s G_i + \hat{\phi}_{\mathbf{x}}\hat{\mathbf{x}}_{\mathbf{i}}}\right) \quad \text{(Eq. 2.1)}$$

1. **No treatment effect and censoring occurs only at the end of the study**.

   The hazard of the event at time T is given using the scale $b_i = d_0(t)e^{\phi_s G_i}$, where two baseline scale parameters $d_0(t) = 15$ and $d_0(t) = 50$ were considered. The baseline scale parameter controls the length of event times acting like a mean for the spread of the simulated distributional values. A larger value for the baseline scale parameter means on average a patient's survival time is longer. Since there was no censoring before the end of the study, the observed event time for the i'th patient was $\tau_i = T_i$ if the event occurred before the end of the study; otherwise $\tau_i = Z$ (i.e. replaced by the end of study time).

2. **Random censoring due to drop-out before the end of the study**.

   The hazard of the event at time T is given using the scale $b_i = d_0(t)e^{\phi_s G_i}$, where a baseline scale parameter of $d_0(t) = 15$ is considered. The censoring time of the i'th individual, c_i, was simulated from an exponential distribution with scale parameters of 20, 40 and 60. The censoring time was assumed to be independent of the SNP. The end of study time is fixed at 40 days; therefore an exponential scale parameter of 20 for censoring times corresponds to approximately 50% censored observations during the study, a scale of 40 corresponds to approximately 30% censored observations, and a scale of 60 corresponds to approximately 20% censored observations. If censoring occurred before the end of the study for the i'th patient, they were assumed to have dropped out at that time, and their observed time $\tau_i = c_i$. If the simulated censoring occurred after the end of the study for the i'th patient, their observed event time was $\tau_i = T_i$ if the event

occurred before the end of the study; otherwise $\tau_i = Z$ (i.e. replaced by the end of study time).

3. **Recruitment period during first ten days of the study and censoring occurs only at the end of the study**.

   The hazard of the event is given using the scale $b_i = d_0(t)e^{\phi_s G_i}$, where a baseline scale of $d_0(t) = 15$ is considered. The recruitment time, r_i, was simulated from a Uniform distribution over the first ten days of the study. Since there was no censoring before the end of the study, the observed event time for the i'th patient was $\tau_i = T_i - r_i$, if the event occurred before the end of the study; otherwise $\tau_i = Z - r_i$.

4. **Multiple treatments with variable effects on outcome and censoring**.

   Patients were randomly assigned to one of four treatments (A, B, C or D). Treatment A increased the hazard of the event at any given time, while treatment C resulted in increased random censoring due to adverse drug reaction before the end of the study. The hazard of the event using the scale is given by $b_i = d_0(t)e^{\phi_s G_i + \hat{\phi}_{\mathbf{x}}\hat{\mathbf{x}}_{\mathbf{i}}}$, where the baseline scale, $d_0(t) = 15$, $\hat{\phi}_{\mathbf{x}} = (0.2, 0, 0, 0)$ is the effect of treatment A, B, C, D, and $\hat{\mathbf{x}}_{\mathbf{i}}$ is an indicator variable taking the value 1 if the i'th patient is assigned to treatment A, and 0 otherwise. If the patient is assigned to treatment C, censoring occurs more frequently before the end of the study, with time, c_i, simulated from an exponential distribution with a scale parameter of 10. All other treatments have a scale parameter of 30 to demonstrate the effects of censoring between treatment C and all other treatments. Similar to Scenario 2, if censoring occurred before the end of the study, the patient was assumed to have dropped out at that time, so that $\tau_i = c_i$. If censoring occurred after the end of the study, the observed event time for the i'th patient was $\tau_i = T_i$ if the event occurred before the end of the study; otherwise $\tau_i = Z$ (i.e. replaced by the end of study time). Under this same design, survival times were simulated with a significant SNP-treatment interaction effect. The hazard now becomes $b_i = d_0(t)e^{\phi_s G_i + \hat{\phi}_{\mathbf{x}}\hat{\mathbf{x}}_{\mathbf{i}} + \phi_\gamma G_i \hat{\mathbf{x}}_{\mathbf{i}}}$ with the interaction effect $\phi_\gamma = 0.2$, and $\hat{\mathbf{x}}_{\mathbf{i}}$ defined as

above.

In parallel, for every scenario above, the event time was dichotomised at the end of the study, Z, such that the binary outcome for the i'th patient, $y_i = 1$ if $T_i < Z$, and 0 otherwise. For scenarios 2 and 4, patients censored before the end of the study are treated as missing for the binary outcome.

2.3.3 Statistical Analysis

For each scenario, the association between the SNP and the TTE was tested under a Cox PH model (Eq. 1.3) and with the binary outcome in a logistic regression framework (Eq. 1.12). For scenarios 1, 2 and 3, the linear component of the regression model included an effect of the SNP only. However, in scenario 4, the linear predictor was extended to include an indicator variable to account for the vector of treatment effects and an interaction effect. The p-value and effect size estimates (log-hazard ratio for Cox PH model and log-odds ratio for logistic regression model) are the output of interest.

2.3.4 Results

The power and type-I error rate was evaluated for each test to detect association of the SNP with the outcome at a 5% significance threshold, approximated by the proportion of replicates for which $p < 0.05$ for the SNP effect.

Figure 2.3 presents the power to detect association of the SNP, under scenario 1, with TTE using a Cox PH model and with the dichotomised outcome at the end of the study in a logistic regression framework. The two plots present power, for the different end of study times, for a baseline scale of 15 (left) and 50 (right). For a baseline scale of 15, the majority of individuals will have experienced an event by day 20, and almost all will have experienced the event by day 60. In this setting, the number of censored observations at the end of the study is low, and the Cox PH model can directly account for the times at which the events occurred. On the other hand, the logistic regression

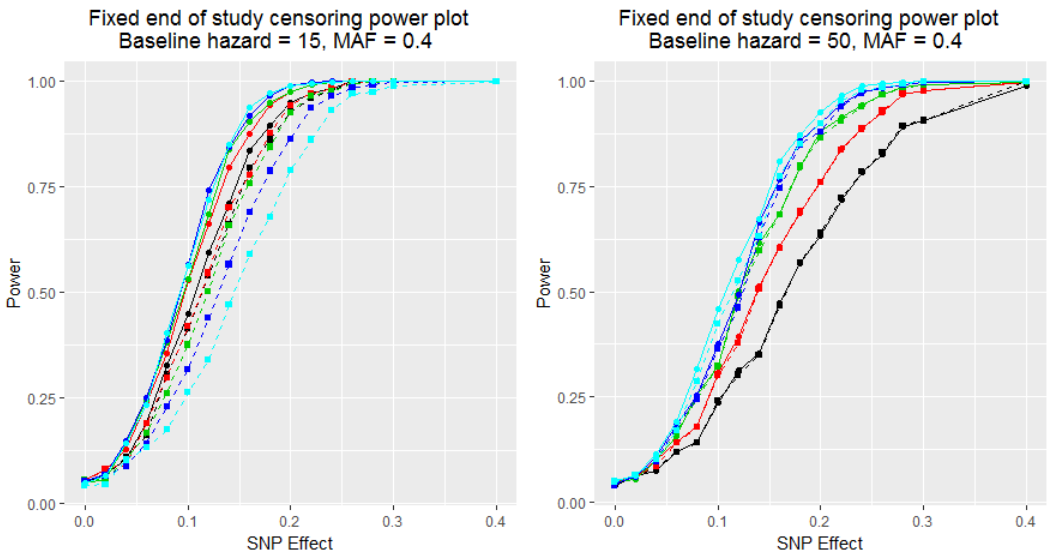


Figure 2.3: Scenario 1 power plots, where the end of study time is varied, and with fixed end of study time censoring. Left plot is the scenario with a baseline scale parameter of 15, and the right plot is for the scenario with a baseline scale parameter of 50. Power is estimated at a 5% significance threshold. Lines with circular points characterise the Cox PH model and lines with square points are the logistic regression model. The colour of the line represents the end of study time: 20 days (black); 30 days (red); 40 days (green); 50 days (blue) and 60 days (cyan).

model loses power as the end of study time increases, because the number of cases and controls becomes more imbalanced. As a result, the difference in power between the two models increases with the end of study time. For a baseline scale of 50, far fewer events occur during the study period. As a consequence, there is less to be gained through direct modelling of event times, and the logistic regression analysis has almost identical power to the Cox PH model. An end of study at 60 days is the most favourable cut-off point for the logistic regression model as it has the greatest balance of events and non-events occurring during the trial.

Figure 2.4 presents a comparison of $-log_{10}$ p-value and estimated effect sizes from scenario 1 for a baseline scale of 15, SNP effect of 0.1 and an end of study time of 40. These parameter settings achieve a power of 50% to detect associations using the Cox PH model. These results highlight that effect sizes from the two approaches are highly correlated, but the signal of association is more often stronger under the Cox PH model than logistic regression model. Figure 2.5 presents the power to detect association of the SNP, under scenario 2, where censoring can occur before the end of the study. As

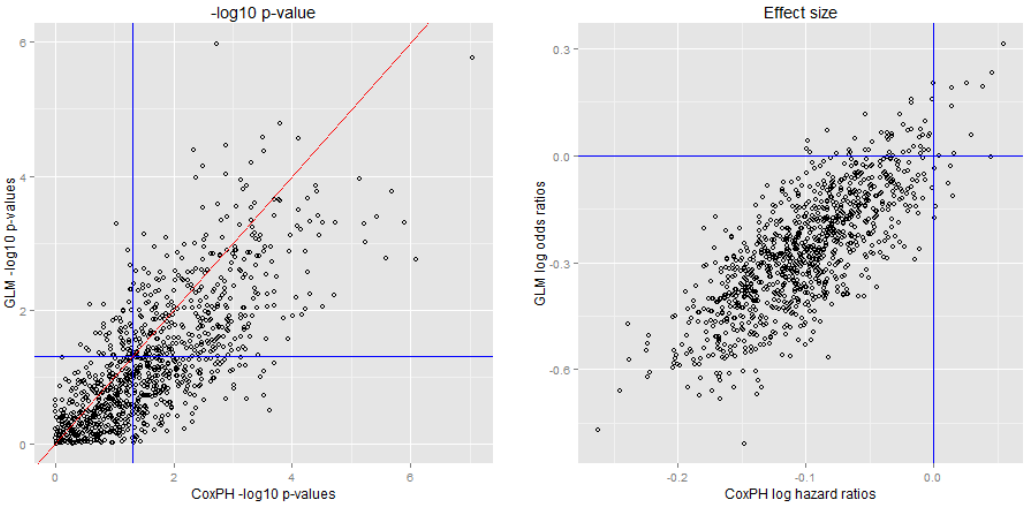


Figure 2.4: Scenario 1 $-log_{10}$ p-value and effect size plots. Blue lines represent 5% significance levels (left) and log-hazard or log-odds ratios of 1 (right). Baseline hazard = 15, MAF = 0.4, SNP effect = 0.1 and End of study = 40. This setting represents a Cox PH model with a 50% power to detect associations between SNPs and survival outcome.

censoring increases, the power of both models decreases. However, the reduction is more dramatic for the logistic regression modelling of dichotomised outcomes because the number of individuals contributing to this analysis decreases, with a consequent reduction in power. Figure 2.6 presents a comparison of $-log_{10}$ p-value and estimated effect sizes from scenario 2 for a baseline scale of 15, SNP effect of 0.1, an exponential censoring scale parameter of 60 and an end of study time of 40. These parameter settings achieve a power of 50% to detect associations using the Cox PH model. These results highlight that effect sizes from the two approaches are again highly correlated, and the signal of association is typically weaker under logistic regression modelling of the dichotomised outcome than the Cox PH model.

Figure 2.7 presents the power to detect association of the SNP and outcome, under scenario 3, which incorporates a variable recruitment period for individuals at the start of the study. The results are similar to those observed for scenario 1 with a baseline scale of 15. As for scenario 1, the difference in power between the two modelling approaches depends on the proportion of individuals in which the event has occurred, and is maximised for an end of study time of 60 days because most individuals will have

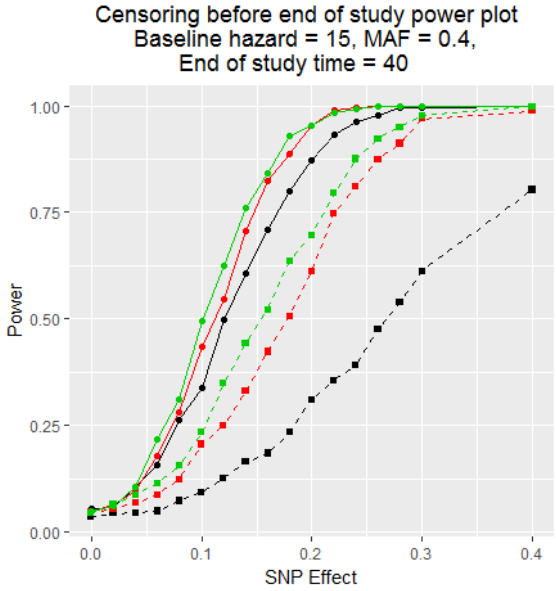


Figure 2.5: Scenario 2 power plot, where the end of study time is fixed at 40 days, but with random censoring during the study period. Power is estimated at a 5% significance threshold. Lines with circular points characterise the Cox PH model and lines with square points the logistic regression model. The colour of the line corresponds to the rate of censoring defined by the scale parameter of the exponential distribution: scale 20 (black); scale 40 (red) and scale 60 (green).

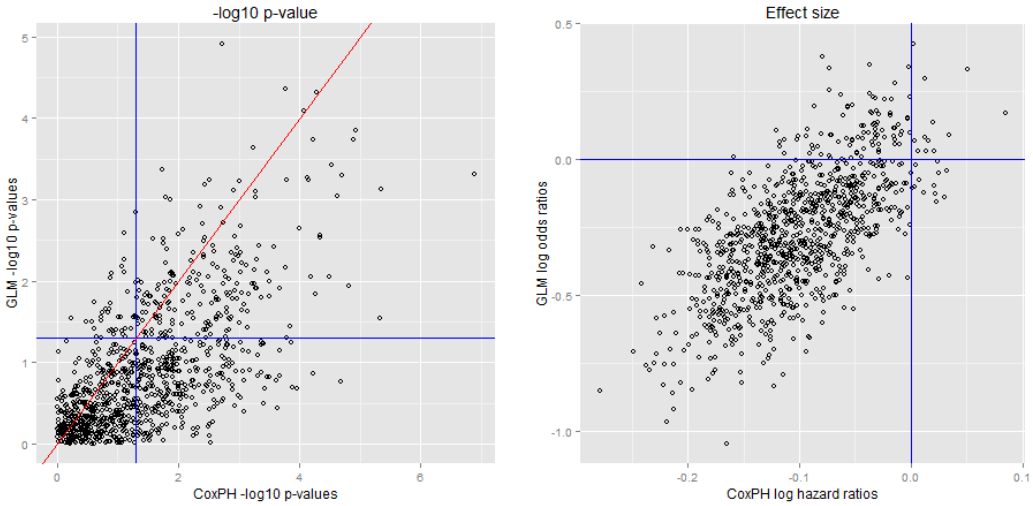


Figure 2.6: Scenario 2 $-log_{10}$ p-value and effect size plots. Blue lines represent 0.05 significance levels (left and log-hazard or log-odds ratios of 1 (right). Baseline hazard = 15, MAF = 0.4, SNP effect = 0.1, Censoring scale = 60 and End of study = 40. This setting represents a Cox PH model with a 50% power to detect associations between SNPs and survival outcome.

experienced the event by this time, even with a variable recruitment period. An end of study of 20 days shows the smallest difference in power between the models because very few events occur with the majority of individuals surviving to the end of the study.

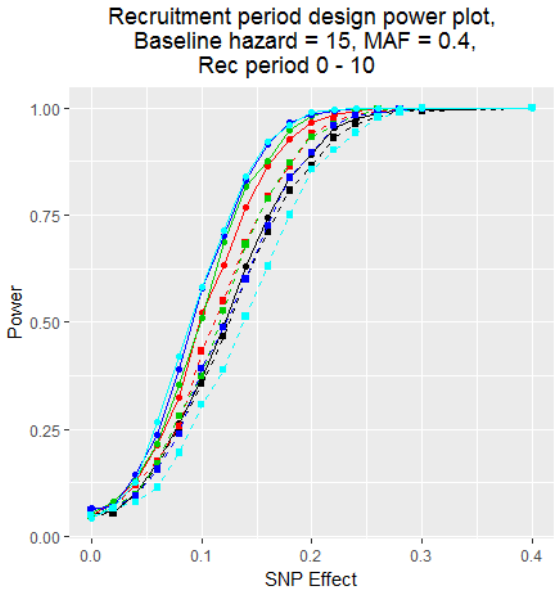


Figure 2.7: Scenario 3 power plot, with variable recruitment period, but with fixed end of study time censoring. Power is estimated at a 5% significance threshold. The recruitment period is between 0 and 10 days. Lines with circular points characterise the Cox PH model and lines with square points the logistic regression model. The colour of the line represents the end of study time: 20 days (black); 30 days (red); 40 days (green); 50 days (blue) and 60 days (cyan).

Figure 2.8 presents a comparison of $-log_{10}$ p-value and estimated effect sizes from scenario 3 for a baseline scale of 15, SNP effect of 0.1 and an end of study time of 40. These parameter settings achieve a power of 50% to detect associations using the Cox PH model. As with previous scenarios, the effect size estimates obtained from the two models are highly correlated, but signals of association are generally stronger under the Cox PH model than the logistic regression model.

Figure 2.9 presents the power of the two modelling approaches for scenario 4, where TTE is dependent on treatment. The first plot (left) includes a treatment effect of 0.2 on survival times with both analysis models adjusting for the treatment covariate. The second plot (right) includes both a treatment effect of 0.2 and a SNP-treatment interaction of 0.2, but incorporates only the treatment covariate in the analysis models (because an interaction effect is not typically assumed, a priori). It is clear from both plots that as the end of study time increases, which results in more censoring during the study period, the power of the logistic regression model is substantially reduced. Introducing a SNP-treatment interaction increases power for both analytical approaches

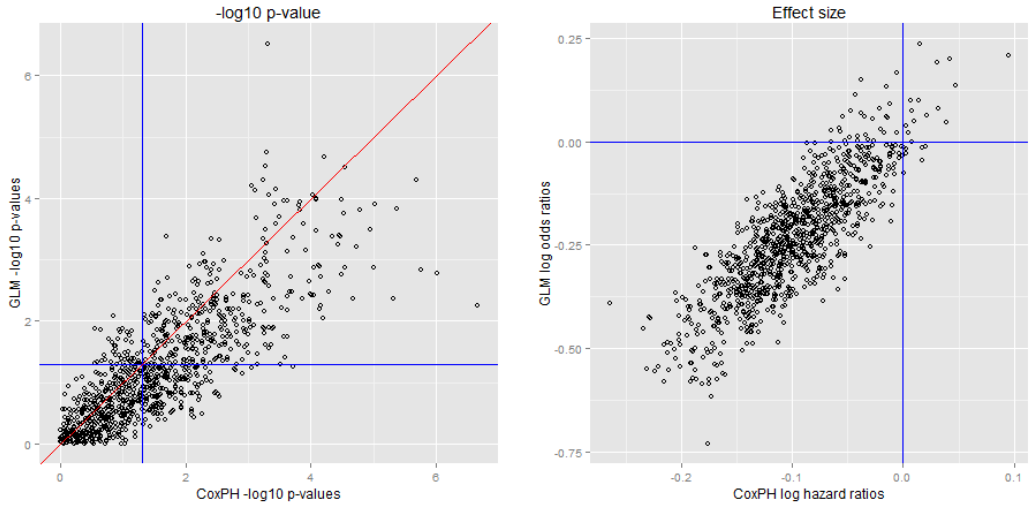


Figure 2.8: Scenario 3 $-log_{10}$ p-value and effect size plots. Blue lines represent 0.05 significance levels (left) and log-hazard or log-odds ratios of 1 (right). Baseline hazard = 15, MAF = 0.4, SNP effect = 0.1 and End of study = 40. This setting represents a Cox PH model with a 50% power to detect associations between SNPs and survival outcome.

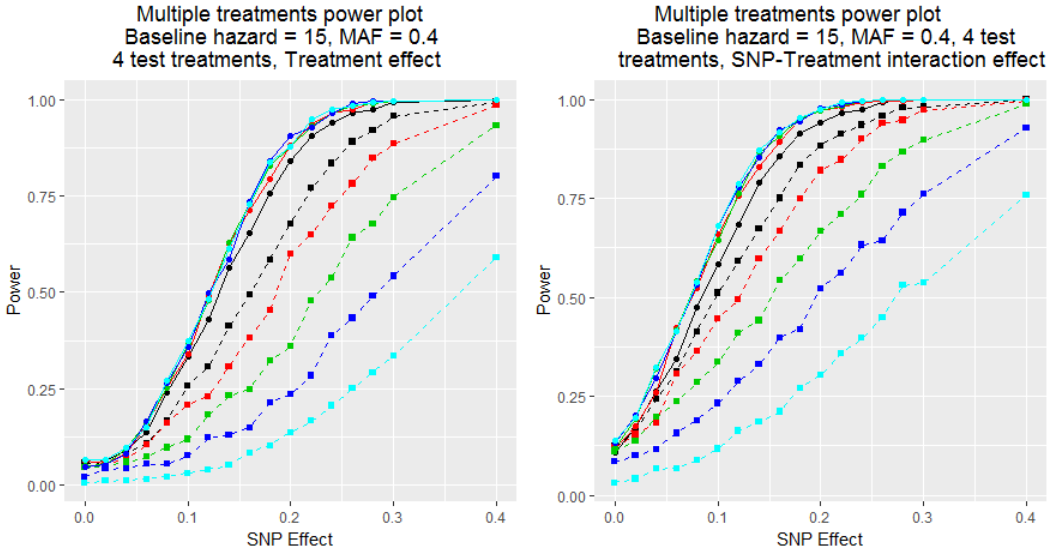


Figure 2.9: Scenario 4 power plots, where the end of study time varied, and there was random censoring during the study period. Power is estimated at a 5% significance threshold. Lines with circular points characterise the Cox PH model and lines with square points the logistic regression model. The colour of the line represents the end of study time: 20 days (black); 30 days (red); 40 days (green); 50 days (blue) and 60 days (cyan).

because the marginal effect of the SNP is increased, even if the interaction effect itself is not taken account of in the analysis model.

Figure 2.10 presents a comparison of $-log_{10}$ p-value and estimated effect sizes from

scenario 4 for a baseline scale of 15, SNP effect of 0.1, treatment effect of 0.2, interaction effect of 0.2 and an end of study time of 40. These parameter settings achieve a power of 50% to detect associations using the Cox PH model. These results highlight that the effect sizes from the two approaches are correlated, but the signal of association is more often stronger under the Cox PH model with very few replicates with more significant associations found towards the logistic regression model.

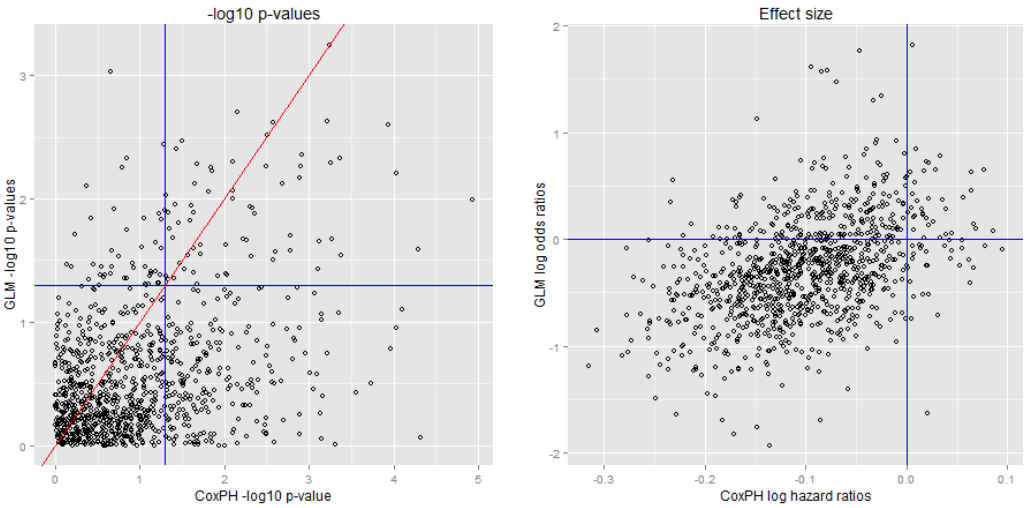


Figure 2.10: Scenario 4 $-log_{10}$ p-value and effect size plots. Blue lines represent 0.05 significance levels (left) and log-hazard or log-odds ratios of 1 (right). Baseline hazard = 15, MAF = 0.4, SNP effect = 0.1, Treatment effect = 0.2, Interaction effect = 0.2 and End of study = 40. This setting represents a Cox PH model with a 50% power to detect associations between SNPs and survival outcome.

For all parameter settings, the Cox PH analysis of TTE was always at least as powerful as the logistic regression model, as expected. Across scenarios, the greatest difference in power occurred when the end of study time was extended because the number of individuals experiencing the event was maximised, and there was more information in the event times themselves. It is also true to say that as the number of censored observations increases the more missing data there is for the logistic regression approach, resulting in a loss in power to detect associations. While the power of the Cox PH model stays relatively constant, that of the logistic regression approach gets substantially weaker because of the increased imbalance in the number of cases and controls.

2.4 Application to the SANAD Study

The Standard And New Anti-epileptic Drugs (SANAD) Study was an unblinded randomised controlled trial in hospital-based outpatient clinics across the UK comparing the effects of various drugs (efficacy) on patients with epilepsy. A sub-study of SANAD conducted by Leschziner et al. (2006) was initiated to investigate the impact of genetic variation on response to anti-epileptic drugs. They considered 503 epilepsy patients receiving one of six anti-epileptic drugs over a follow-up period of between 84 and 2296 days. The study evaluated the evidence of association of 501 SNPs mapping to/near the *ABCB1* gene with time to 12-month remission, time to first seizure, and time to drug withdrawal due to inadequate seizure control or adverse reactions.

For the purpose of this example, the original dataset was manipulated to illustrate similar designs to that of the simulation study, incorporating binary outcomes as an end of study time which is not specified in the original trial. The purpose of this analysis was not to replicate the initial findings of the study but to compare results of modelling TTE as survival or dichotomous outcomes. Specifically, testing for association of SNPs with time to first seizure within the first 12, 24, or 36 months of follow-up using a Cox PH model, and with a dichotomised outcome, at the same follow-up time points, in a logistic regression model. The numbers of missing observations due to censoring before the end of the study, for the logistic regression model at 12, 24 and 36 months, were 56, 69 and 95, of the 503 patients, respectively. Before analysis, SNPs were eliminated based on a MAF less than 1% and missing genotype rate greater than 5%. For a full breakdown of summary statistics refer to Table 2.2.

To calculate the observed time for each patient and the corresponding outcome (survival and binary) based on the three new end of study times, a combination of the original outcomes of time to drug withdrawal and time to first seizure was used. Figure 2.11 shows a detailed diagram of how each outcome was developed. Adjustment for covariates such as treatment was not undertaken in this illustrative analysis.

Sample size						Associated SNPs (p <0.05, MAF >1% & Missing <5%)			
Outcome	Censored	Censored before end of study	Events	Non-events	SNPs after QC	Total	Cox PH	GLM	Shared
Time to Seizure 12 months	156	56	347	100	116	10	6	5	1
Time to Seizure 24 month	138	69	365	69	116	16	9	8	1
Time to Seizure 36 months	132	95	371	37	116	10	9	1	0

Table 2.2: Summary statistics from SANAD study. Missing refers to the percentage of missing genotype data. Shared refers to the number of shared associations discovered between the Cox PH and logistic regression models. Abbreviations: SNPs, single nucleotide polymorphisms; GLM, generalised linear model (logistic regression); PH, proportional hazards; MAF, minor allele frequency; QC, quality control. p is the value of significance.

Patient	Seizure time	Withdrawal time	Censoring (Withdrawal)	12 month EOS	24 Month EOS	36 Month EOS
1	686	686	0	365	730	1095
2	95	13	1	365	730	1095
3	36	2228	0	365	730	1095
4	495	1083	0	365	730	1095

Patient 4 survives to the end of the 12 month study (seizure time and withdrawal time are both after this point) therefore the observed time is 365 and they are censored (survival) and non-event (binary). For 24 and 36 month end of study the patient has the event of first seizure.

Patient 2 - time of seizure occurs at 95 days however they are censored due to drug withdrawal occurring at 13 days.

Patient 3 - time of seizure occurs at 36 days before the drug withdrawal and end of study times for all 3 studies therefore the event of seizure has occurred uniformly across all outcomes.

Patient 1 - Seizure and drug withdrawal occur at the same time before the end of study therefore the patient is censored. For 12 month follow-up the observed time is 365 days indicating the patient has survived to the end.

Observed Time (12)	Time-to-event Outcome (12)	Binary Outcome (12)
365	0	0
13	0	NA
36	1	1
365	0	0

Observed Time (24)	Time-to-event Outcome (24)	Binary Outcome (24)
686	0	NA
13	0	NA
36	1	1
495	1	1

Observed Time (36)	Time-to-event Outcome (36)	Binary Outcome (36)
686	0	NA
13	0	NA
36	1	1
495	1	1

Figure 2.11: Outcome calculation diagram for SANAD study sample dataset. Abbreviation: EOS, end of the study.

2.4.1 Results

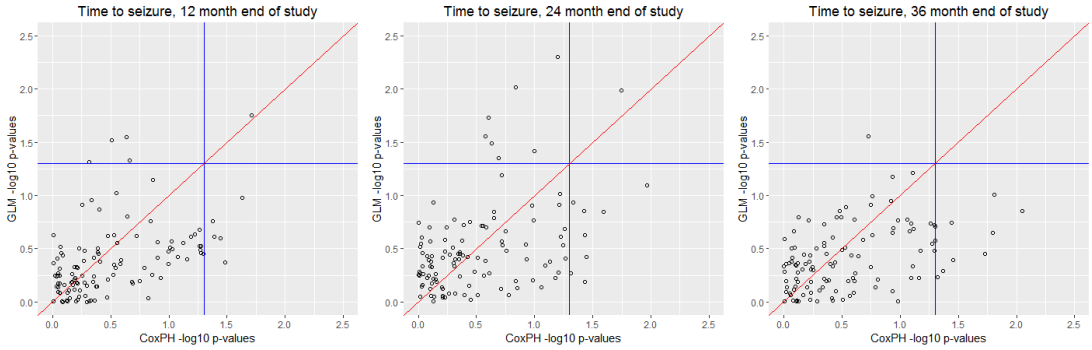


Figure 2.12: $-log_{10}$ p-value plots for the outcomes time to seizure at 12 months (left), time to seizure at 24 months (middle) and time to seizure 36 months (right). Blue lines represent 0.05 significance levels.

Figure 2.12 depicts the $-log_{10}$ p-value for association of SNPs with time to first seizure from the SANAD study. Each point corresponds to a SNP, with the p-value for the SNP effect from the Cox PH model on the x-axis and the p-value for the SNP effect from the logistic regression model on the y-axis. The two blue lines indicate a nominal 5% significance threshold for the association. The top left quadrant indicates significant SNPs found only by the logistic regression model, the bottom right quadrant shows significant SNPs found only by the Cox PH model, and the top right quadrant indicates

significant SNPs found by both models. The key observation from this analysis is that, as the study period increases, the number of associated SNPs found by the logistic regression model is reduced. This result supports the findings of the simulation study because, as the end of study time increases, there are more censored events within the sample. For the outcomes of time to seizure at 12 and 24 months, there is little difference between significant SNPs found by the logistic regression or Cox PH models. However, for an end of study time of 36 months, the logistic regression model is able to detect only one association.

2.5 Discussion

As expected, there was very little to gain by simplifying the TTE outcome to binary and using sub-optimal methodology. The Cox PH model was demonstrated to be uniformly more powerful than the logistic regression analysis of dichotomised outcomes across simulation scenarios and generated stronger signals of association in the SANAD study. However, the difference in power between methods was highly dependent on the rate of censoring and number of events occurring within the study period. If few events occur and the majority of individuals survive until the end of the study, then there is little to be gained by modelling TTE outcomes. In these scenarios, a recommendation would be to perform an initial evaluation of SNP association signals using computationally efficient software with dichotomised outcomes identifying SNPs for further evaluation with more rigorous, and computationally demanding, survival modelling, thereby providing an effective screening tool. Consequently, this has important implications for the development of analytical protocols in pharmacogenetic studies.

The power of the logistic regression model is significantly affected by the two types of censoring: (i) at the end of study because the event has yet to occur; and (ii) during the study period because of the occurrence of an adverse event or drop-out. If there is a lot of censoring during the study, these observations are treated as missing in the logistic regression model and therefore result in a lack of power to detect associations between

SNPs and outcome. The rate of censoring at the end of the study (i.e. the number of events that have occurred) will impact the ratio of cases to controls in the dichotomised outcome. For a given sample size, imbalanced studies will have less power than those for which the number of cases and controls is equal.

Nearly all of the TTE articles reviewed applied the Cox PH model without giving an indication of testing the PH assumption. Gregers et al. (2015) is one of the few articles that were reviewed that justified the use of a Cox PH model. To assess the proportionality assumption they used Schoenfeld (Schoenfeld 1982) and Martingale residuals (Therneau et al. 1990). Model checking is essential and can help identify the correct choice for analysis. For the majority of studies, survival modelling has been condensed into using Kaplan-Meier curves and log-rank tests to explain differences between groups and the Cox PH model to quantify the difference using hazard ratios. This in itself is a big limitation because other statistical models may be more applicable to the data.

The SANAD study results were consistent with the findings of our simulation study. The candidate gene study was followed up with a GWAS several years later (Speed et al. 2014). This study was a multi-centre study of two cohorts of 916 newly treated epilepsy patients. The clinical outcome of interest for this larger study was 12-month remission. Patients achieving 12-month remission from seizures were defined as "responders", and patients failing to achieve 12-month remission were defined as "non-responders". Patients followed for less than 12 months were excluded from the study. For this study the association analyses were performed using a logistic regression model in PLINK, thereby losing all time related patient information. Doing this again emphasises the need for flexible software for the analysis of TTE data that can efficiently handle the scale and complexity of genetic data throughout the genome.

A final observation from the literature review was that very few studies undertook sample size and power calculation before study engagement. Yip et al. (2014) a candidate gene study determining associations between SNPs and treatment response,

is one study which outlines a statistical power calculation procedure. This is an essential step that will be investigated further in the next chapter.

CHAPTER 3

DATA SIMULATION AND POWER CALCULATION

3.1 Overview

In the previous chapter, a few key observations were made regarding power calculations and genetic data simulation. The first observation was the absence of sample size calculation protocols in published papers within pharmacogenetics. This observation is informative since power calculators are currently available for the design of genetic association studies of binary phenotypes and quantitative traits, but not for time-to-event (TTE) outcomes, which are of particular relevance in pharmacogenetics. With the rapid emergence of pharmacogenetic association studies of single nucleotide polymorphisms (SNPs) and the complexity of the clinical outcomes they consider, there is a definite need for software to perform power calculations of TTE data over a range of design scenarios and analytical methodologies. The second observation was the lack of motivation for the choice of analysis methodology. Most TTE studies within pharmacogenetics stated use of the Cox proportional hazards (PH) model without considering alternatives or testing the PH assumption.

3.1.1 Objectives

The objective of this chapter is to define software to perform power calculations for genetic association studies of TTE outcomes that consider a range of design scenarios and analytical approaches. The simulation machinery developed in Chapter 2 has been utilised in the development of algorithms into the simulation and power calculation software design.

3.2 Simulating Realistic Genetic Data

One approach to assess power requires simulation of genotype and outcome data. In Section 2.3, scenarios were designed based on a review of the literature and simulated data in the R (R Core Team 2013) statistical environment using generated random deviates from statistical distributions. However, there is specific software used for simulating genetic data, such as GPOPSIM (Zhang et al. 2015), a simulation tool for pedigree, phenotypes, and genomic data or genomeSIM (Dudek et al. 2006) for the simulation of large-scale genomic data in population-based case-control samples. HAPGEN2 (Su et al. 2011) is software that uses a simulation-based algorithm based on a re-sampling method. It simulates realistic SNP data based on known haplotypes from a reference panel, for instance, the 1000 Genomes Project (Auton et al. 2015) or HapMap3 (Altshuler et al. 2010). HAPGEN2 generates genotype and phenotype files for case-control studies ready to be used by analysis software.

Software to perform data simulation for GWAS with TTE studies are non-existent. Simulations can be undertaken separately for genotype and outcome, however incorporating genetic effects on event times is difficult to achieve without bespoke software.

3.3 Importance of Power Calculations and Sample Size

Power and sample size calculations are an essential component of study design. They inform us about the required sample size to detect a desired effect size with sufficient power at a given level of significance. These calculations can be conducted through simulating many replicates of data based on the investigators' input parameters or previous pilot data. These replicates are then analysed using a statistical test to produce results on power, false-discovery rate and other metrics.

Low et al. (2014) states that for "underpowered studies, and large heterogeneity of study designs, collaborative efforts are needed to validate these findings and overcome the limitations of GWA studies before clinical implementation". This sentence defines how

necessary power calculations are to all types of studies because without it investigators would be making assumptions through uninformative speculation.

Tools are readily available for GWAS of binary phenotypes and quantitative traits. These include the freely available Genetic Power Calculator, developed by Purcell et al. (2003), which was an innovative web-based platform produced in 2001, for performing power calculations for the design of linkage and association genetic mapping studies of complex traits. It offers users options for discrete and quantitative trait power calculations under both case-control and family-based association. Software such as CaTS (Skol et al. 2006) followed this, introducing a two-stage GWAS power calculator, which allows for a discovery and replication phase. CaTS jointly analyses data from the first stage (the proportion of the available samples genotyped on a large number of markers), and the second stage (the portion of these markers that are later followed up by genotyping them on the remaining samples). Most recently, GAS (Johnson 2017), a user-friendly web application for case-control genetic association power calculation, is at the forefront of the development of interactive applications that are accessible to everyone without download. However, as yet, software is not available to determine adequate sample size for pharmacogenetic GWAS of TTE outcomes. This would be useful in many situations, for instance, where the impact of alternative treatments, and potentially their interaction with SNPs is often of relevance in the study design protocol.

Owzar et al. (2012) presents methodology and simulation study results for asymptotic and empirical power and sample size calculations for SNP association studies with TTE outcomes. They also provide an R package '*survSNP*' which facilitate the calculations in the paper using a Cox score test, implemented using the '*survival*' package developed by Therneau (2015). However, the package does not allow users the flexibility to calculate the power and sample size based on different statistical models and study designs. A graphical user interface (GUI) based software would also be easier to use for those with limited R experience.

The lack of data simulation and power calculation tools available for pharmacogenetic

TTE studies has prompted the development of the tool SurvivalGWAS_Power. The design of the software draws influence from the review of current software in the previous section. As a result, the software has a user-friendly interface and attempts to present the results output in the most informative way possible, through plots and statistical metrics such as the power based on sample size, effect size and allele frequency.

3.4 SurvivalGWAS_Power

SurvivalGWAS_Power was developed from a Windows form based calculator application that performed basic arithmetic in C++. The program was then ported over to C# which, provided more flexibility in the design of the GUI and offered a vast catalogue of .NET libraries to utilise. Even though it was initially designed for pharmacogenetic studies based on the literature review in Chapter 2, the potential for application more widely to other study designs is available.

3.4.1 Implementation

SurvivalGWAS_Power was built as a Windows application, utilising pre-designed frameworks Math.NET (https://www.mathdotnet.com/) and Accord.NET (Souza 2014), for the generation of pharmacogenetic data and statistical analyses, respectively. The software was compiled on a Windows operating system (O/S) using the integrated development environment (IDE) Visual Studio 2013 (https://www.visualstudio.com/). SurvivalGWAS_Power requires specification of genetic parameters, such as the magnitude of the SNP effect on the outcome and the effect allele frequency (EAF). The varied collection of design scenarios includes adding a recruitment period, SNP-treatment interactions, and different censoring options (for example, withdrawal due to an adverse treatment event). The scenarios were created using the thorough examination of published pharmacogenetic studies in the literature review of Chapter 2. The power calculations are performed by simulating multiple datasets based on the user-specified parameter settings and study design options, specifically testing for SNP associations

(and SNP-treatment interactions, if required) with the TTE outcome across all simulated datasets.

3.4.2 User Interface

As mentioned earlier, the software boasts a GUI, making it understandable and practical to navigate. The main window consists of two panels, the first for design, analysis and parameter inputs, and the second for all output (see Figure 3.1). The menu bar has a "Save Sample Data" option, as well as another choice to store all the datasets from every simulation run. This option might be useful for those who want to test power for methods not supported by the program. The data are saved as a text file, in R statistical software readable format. The interface has been designed to be user-friendly; there are various help buttons to navigate the user through the program in the form of tooltips, and an example of a commonly used pharmacogenetic study design is available as a guide. The inputs are split into two sections: (i) data generation inputs; and (ii) statistical analysis inputs. The user-defined parameter inputs are submitted in text boxes. For a full description of inputs and results see Table 3.1.

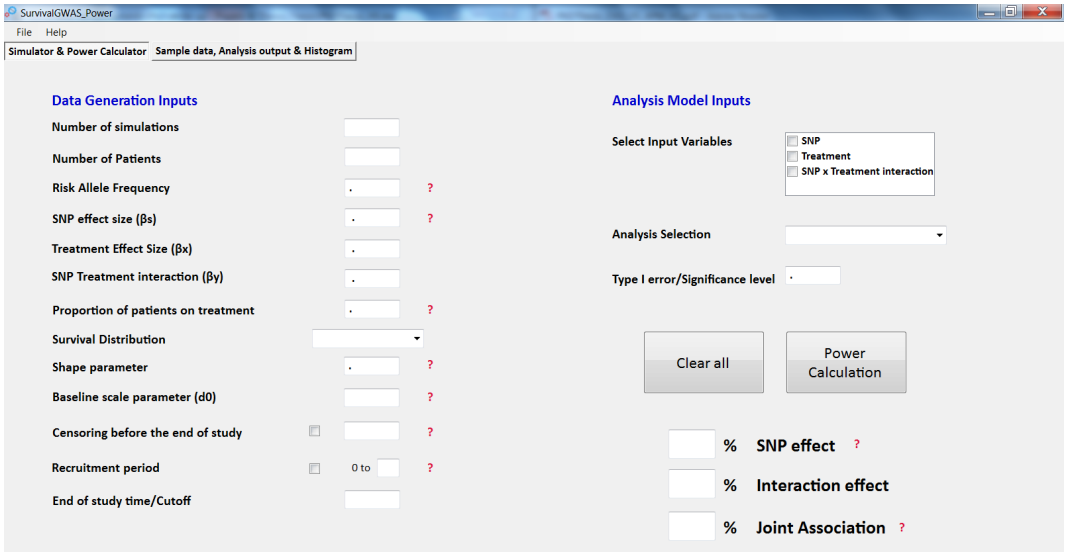


Figure 3.1: SurvivalGWAS_Power v1.5 (Date: 30/11/2017) User interface screen-shot of the simulator and power calculator tab.

Figure 3.1, presents a screen-shot of the front end of SurvivalGWAS_Power. If the user

clicks on the tab at the top labelled "Sample data, Analysis output & Histograms" they will see the additional output from all the analyses and data simulations. Figure 3.2 is a screen-shot of the interface for the results tab. Data nor power has been simulated here, so all four boxes are empty. Output details are discussed using an example in Section 3.6.

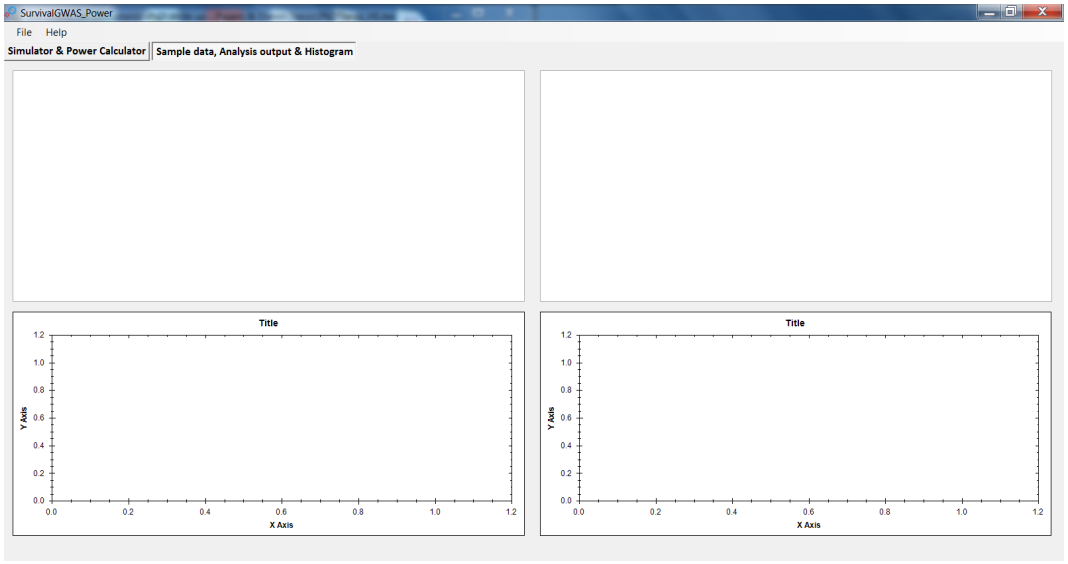


Figure 3.2: SurvivalGWAS_Power v1.5 (Date: 30/11/2017) User interface screen-shot of results tab.

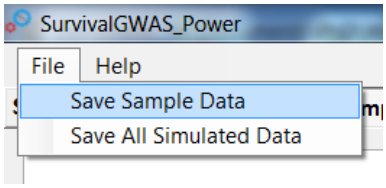


Figure 3.3: File menu of SurvivalGWAS_Power v1.5 (Date: 30/11/2017). Options are to either save the sample data or save data from all simulations.

If the user wishes to save the sample data or data from every simulation run, then this can be done by clicking on the "File" button located on the top menu and then selecting the appropriate choice from the drop-down menu, as shown in Figure 3.3. The reason for this feature was to allow users to use the data for a simulation study or use within other software such as R using methods not supported by the current implementation of SurvivalGWAS_Power. The "Help" button depicted in Figure 3.4, has four options for users to learn more about navigating through the software and details about using and

distributing the software. Figure 3.5 shows a screen-shot of the "About" option from the "Help" menu, which displays the version and general description of the software.

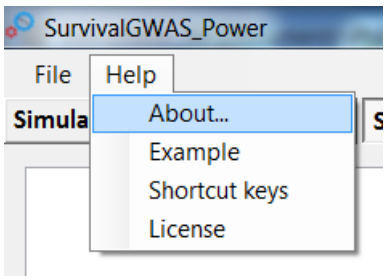


Figure 3.4: Help menu of SurvivalGWAS_Power v1.5 (Date: 30/11/2017). Options include launching an About prompt, example, program shortcut keys and the software license.

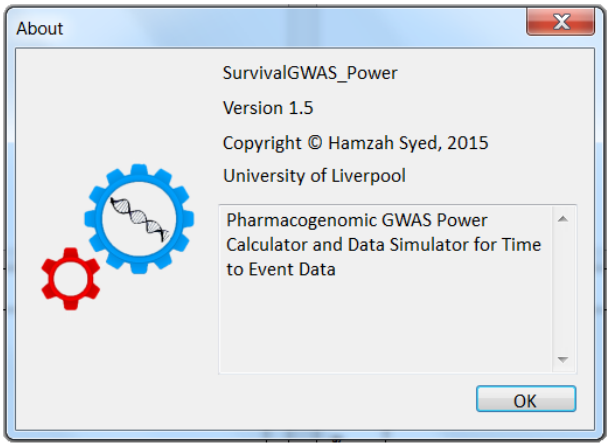


Figure 3.5: About prompt of SurvivalGWAS_Power v1.5 (Date: 30/11/2017). About prompt includes copyright information, logo, version and software description.

Data generation inputs	
Number of simulations	The number of simulated datasets.
Number of patients	The number of patients within each simulated dataset.
Effect allele frequency	Each SNP is simulated using a binomial distribution. Genotype AA=0, AB=1, BB=2. A value of 0.4 means that ≈16% of the patients in each dataset will have the genotype BB.
SNP effect size	The log-hazard ratio for each copy of the effect allele relative to the non-effect allele.
Treatment effect size	The log-hazard ratio between treatment and a placebo.

Table 3.1 continued from previous page

SNP-Treatment interaction	The effect size of the interaction between the SNP and treatment.
Proportion of patients on treatment	Placebo = 0, Test Treatment = 1. A value of 0.5 means both treatments are divided equally between the numbers of patients in each dataset.
Survival distribution	This is the Weibull statistical distribution used to simulate each individuals event time. The scale parameter incorporates SNP, treatment and interaction effect sizes and a baseline scale parameter.
Shape parameter	The shape parameter for the Weibull distribution which is used to simulate survival times for each patient. A value < 1 indicates that the failure rate decreases over time. A value = 1 indicates that the failure rate is constant over time, reduces to an exponential distribution. A value > 1 indicates that the failure rate increases with time.
Baseline scale parameter	Determines time scale of patient's survival time. A value of 20 will simulate survival times around 20.
Censoring	Censoring time for each patient. Input for scale parameter of Weibull distribution, with shape parameter = 1. If censoring time <survival time, then the patient is censored with their observed time = censoring time.
Recruitment period	Patient's recruitment time will be between 0 and value entered. Will effect censoring and observed survival time for each patient.
End of study time	The end of study time.
Analysis model inputs	
Input variables	Check the box/boxes of the model terms to include in analysis. SNP has to be selected.

Table 3.1 continued from previous page

Analysis selection	Choice of either the Cox PH model or Weibull regression model.
Significance level	Significance threshold for p-value from analysis.
Power output	
SNP effect %	The number of simulations at which the p-value for the SNP is significant at the threshold value.
Interaction effect %	The number of simulations at which the p-value for the interaction is significant at the threshold value.
Joint Association %	The number of simulations at which the p-value for thejoint association through LRT is significant at the threshold value.

Table 3.1: SurvivalGWAS_Power inputs and results definitions.

3.4.3 Data Simulation Settings

The user will first specify the number of simulations to generate, and the sample size for each of those simulations. For each replicate of data, a SNP genotype (coded as 0, 1 or 2 according to the number of effect alleles) is generated for each individual from a binomial distribution dependent on the EAF, assuming Hardy-Weinberg equilibrium. The user is given the option of incorporating an active treatment against a placebo. Treatment allocation is simulated using a Bernoulli distribution.

TTE for each individual is then simulated on the basis of specified model parameters from a Weibull distribution given their allocated treatment and genotype, which allows for the possibility of a deviation from a PH assumption. The user specifies the value of the shape parameter, a, of the Weibull distribution. A value of $a < 1$ indicates that the failure rate decreases over time. A value of $a = 1$ indicates that the failure rate is constant over time, resulting in PH. A value of $a > 1$ indicates that the failure rate increases with time. The scale parameter of the Weibull distribution is parametrised to

incorporate SNP, treatment, and SNP-treatment interaction effects in generating TTE for each individual. Specifically, the scale parameter for the i'th individual is given by, $b_i = d_0 e^{-\phi_s G_i - \phi_x x_i - \phi_\gamma G_i x_i}$, where G_i is the SNP genotype coded under an additive model for the minor allele, and x_i is the treatment covariate (coded as 0/1 for placebo/active). The value of the baseline scale parameter d_0, is specified by the user. Larger values of d_0 will simulate larger event times; however, this is dependent on the time-scale (days, months, years) the investigator wants to design the trial around. The parameters ϕ_s and ϕ_x are the effect on log-hazard of the effect allele at the SNP, and the treatment effect, respectively, and ϕ_γ is the interaction effect between the SNP and treatment. The user specifies the values of each of these parameters.

By using the same notation as the study design parameters outlined in Section 2.3, the simulated observed TTE outcome can be generated for the following possible scenarios, each of which includes the option of incorporating treatment and SNP-treatment interaction effects. In all scenarios, the simulated TTE of the i'th individual is denoted T_i, and the observed event time (after right censoring) is denoted as τ_i.

1. **Scenario 1 - End of study censoring.**

   This scenario is designed based on a user-specified fixed end of study time, Z. If the event occurs before the end of the study, the observed event time for the i'th individual is $\tau_i = T_i$; otherwise $\tau_i = Z$.

2. **Scenario 2 - Censoring during the study period and at the end of the study.**

   The censoring time of the i'th individual, c_i, is simulated from a Weibull distribution with a user-defined scale parameter and a fixed shape parameter of 1, to illustrate the censoring is constant over time. Small values of the scale parameter will generate more censored observations. If censoring occurs before the end of the study, the individual is assumed to have dropped out at that time, thus $\tau_i = c_i$. If censoring occurs after the end of the study, yet the event occurred before the end of the study, then the observed event time for the i'th individual is $\tau_i = T_i$; otherwise $\tau_i = Z$.

3. **Scenario 3 - Recruitment period and end of study censoring.**

   The recruitment time, r_i, is simulated from a discrete uniform distribution between 0 and a specified end time. There is no censoring before the end of the study and if the event occurs before the end of the study the observed event time for the i'th individual is $\tau_i = T_i - r_i$; otherwise $\tau_i = Z - r_i$.

4. **Scenario 4 - Censoring during the study period and at the end of the study with a recruitment period.**

   The censoring time of the i'th individual, c_i, is simulated from a Weibull distribution with a user-defined scale parameter and a fixed shape parameter of 1. If censoring does not occur before the end of the study, and an event has occurred, an individual will have observed time $\tau_i = T_i - r_i$, unless censored at the end of the study, then $\tau_i = Z - r_i$. If censoring does occur during the study period, then $\tau_i = c_i - r_i$.

In a real study, there are aspects of the design the investigator can control such as the number of patients, inclusion of covariates and statistical methods. However, some features cannot be controlled such as the SNP effect size and allele frequency. Nevertheless, providing these options allows users to prepare for a variety of possible scenarios.

3.4.4 Methodology

As explained in Chapter 1 and from the pharmacogenetic literature review in Chapter 2, there are a variety of models for analysing TTE data. Therefore, SurvivalGWAS_Power gives users the option of testing the power using either a Cox PH model or a Weibull regression model. Providing a choice will allow investigators to explore an alternative option to the Cox PH model. Users can select between running their choice of analysis by fitting a model including (i) the SNP alone; (ii) the SNP and treatment; or (iii) the SNP, treatment and SNP-treatment interaction.

The framework 'Accord.NET' has a built-in Cox PH function, which calculates the partial likelihood and obtains parameter estimates and Wald test p-values (see Eq. 1.10). The Weibull regression model is not supported in this framework, so maximum likelihood estimates of model parameters are obtained using an iterative Newton-Raphson method after maximising the right censored Weibull regression likelihood function (see Eq. 1.7). An issue with the Weibull regression model parameter estimation using the Newton-Raphson method is that it is sensitive to starting values. This sensitivity can lead to unreasonable updates for the shape parameter. Thus a simple exponential model is first run on the data, and the updates of each iteration are very gradual to achieve more accurate coefficient estimates.

3.4.5 Validation

The user interface implements a validation system to track user errors at an input. As the user inputs values into the parameter text boxes, the error provider will check that the entered values are valid, indicating with either a green tick icon on the right of the box or an exclamation mark as a warning for an incorrect entry. Each textbox also has a character limit, and users will be unable to enter non-numerical characters. Before the power calculation begins, the error provider will check that all required information has been entered for a selected scenario. If not, the program will display a prompt notifying the user what the problem is. For example, the user cannot select treatment as an analysis covariate if a treatment effect has not been included in the simulation model.

3.4.6 Output

The output comprises of a sample dataset, a table of the analysis output for each simulation run and two histograms of parameter estimates across simulations: (i) coefficient values for the SNP effect from the regression model; and (ii) $-log_{10}$ Wald p – values for the SNP effect. All histograms can be saved by right-clicking the graph

and selecting "save as image". Power, at the specified significance threshold, α, is approximated by the proportion of replicates for which $p < \alpha$ for the SNP effect on the outcome. Power, at the same significance threshold, is also calculated for the SNP-treatment interaction effect, if this term is included in the analysis model.

The joint association power is the power to detect an association between the SNP adjusting for treatment and SNP-treatment interaction effects with the TTE outcome. The joint association is calculated from a likelihood ratio test (LRT) (see Eq. 3.1) between two models when the interaction check box is selected. The first model includes the SNP, treatment and interaction terms in the log-likelihood function of either the Cox PH (Eq. 1.4) or Weibull regression models (Eq. 1.7) whereas the null model includes the treatment covariate only. This is a 2 degree of freedom χ^2 test. Figure 3.6 illustrates the simplicity of using SurvivalGWAS_Power through a work-flow diagram. All .NET libraries are pre-compiled within the software. The user can interrupt the power calculation process by clicking the "cancel" button at any time.

$$2(\ell(\beta_G, \beta_x, \beta_\gamma) - \ell(\beta_x)) \qquad \text{(Eq. 3.1)}$$

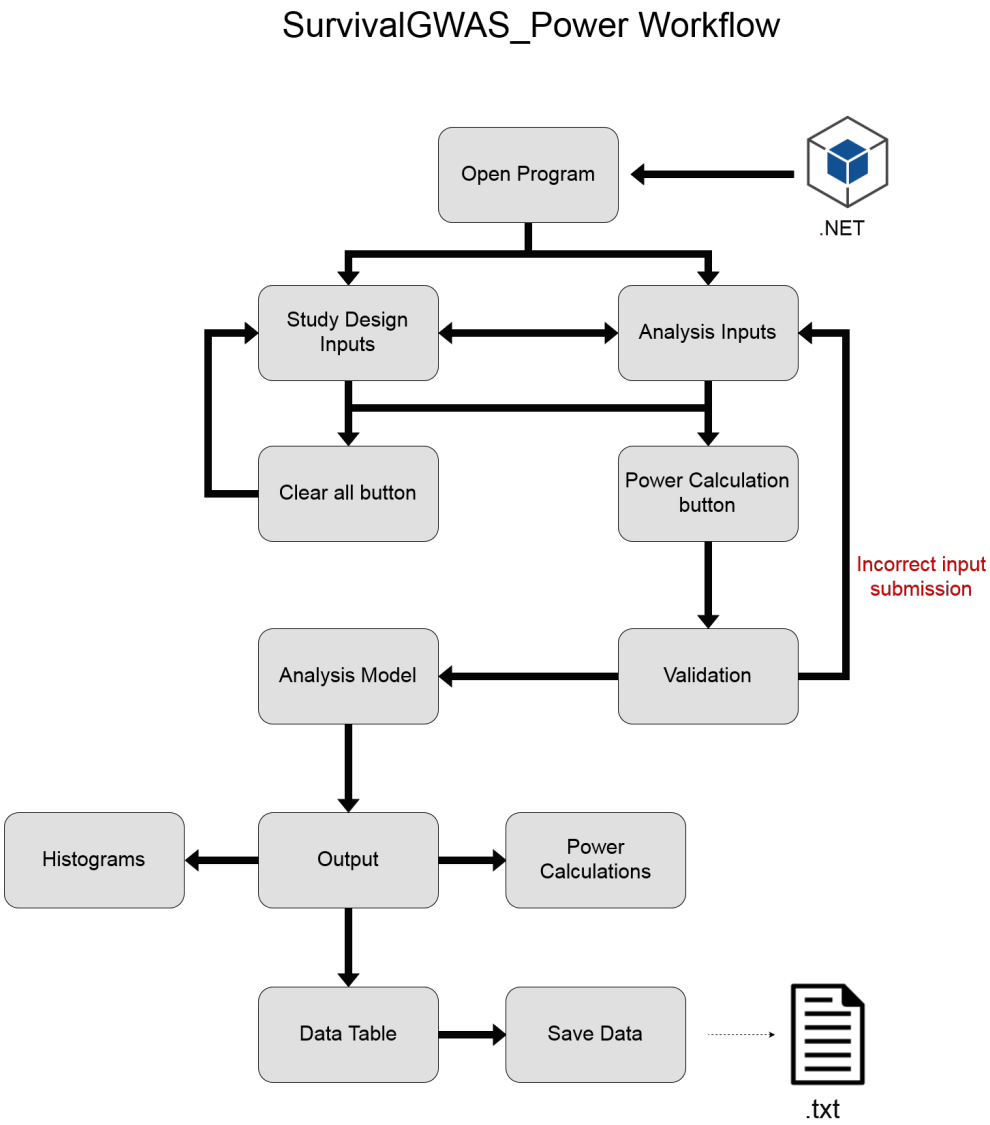


Figure 3.6: Flowchart of SurvivalGWAS_Power v1.5, from data generation to output.

3.4.7 Installation Guide

The software can be downloaded from the University of Liverpool, Statistical Genetics and Pharmacogenomics Research Group software page: https://www.liverpool.ac.uk/translational-medicine/research/statistical-genetics/survival-gwas/. As well as the download, there is a full description of the software. The user must click on the download SurvivalGWAS_Power link and follow the instructions on the screen. Once installed the program should be located in C:/Program Files (x86)/University of Liverpool/SurvivalGWAS_Power. To create a desktop shortcut or pin the application to the taskbar, the user should right-click on the .exe file and select the appropriate option

listed. To open the program, double-click on the .exe file.

3.5 Performance Results

Figure 3.7 presents run times of SurvivalGWAS_Power under two different analyses as a function of the number of simulations: (i) SNP effect only; and (ii) SNP effect, treatment effect and SNP-treatment interaction. Results are presented for the Cox PH and Weibull regression models, for a sample size of 1000 individuals under design scenario 2, outlined in Section 3.4.3.

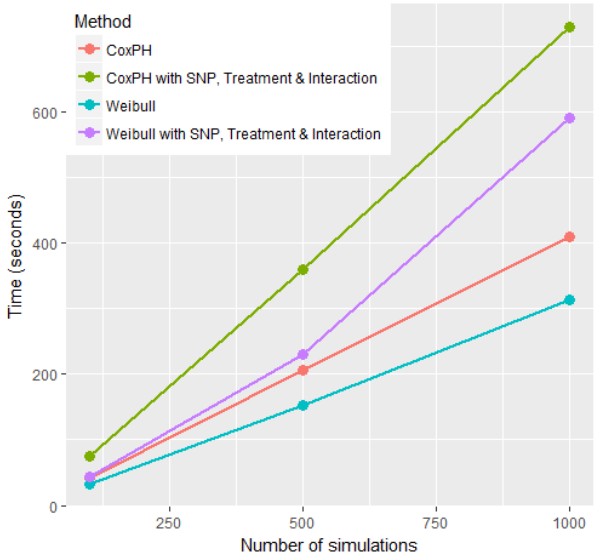


Figure 3.7: Performance of SurvivalGWAS_Power v1.5 assessed by comparing alternative regression models. A sample size of 1000 used for each simulation. All lines represent the terms adjusted for in each statistical model.

Figure 3.7 shows us that under these conditions the Weibull regression model is noticeably faster at processing data than the Cox PH model. However, the speed of the analyses is not informative about the suitability of the models and should not be taken into consideration when designing a study. What it does highlight is that overall the software maintains efficiency, even with the adjustment for treatment and interaction terms.

3.6 Example

SurvivalGWAS_Power has been developed to simulate a large number of datasets to enable efficient estimation of power based on specified model parameters and design scenarios. This section presents the results of example power calculations for a scenario to demonstrate the utility of the software. The example specifically compares two statistical models, while adjustments are made with and without model covariates.

For these examples, it is of interest to investigate the power to detect associations of a pharmacogenetic TTE study. The outcome of interest is known to have a steady increase in risk over the first few months after diagnoses with a very high event rate occurring shortly after. This hazard function is monotonically increasing over a short period. A Weibull distribution with shape parameter 2 and scale 18 is used to simulate this. The patients are being treated with either an active treatment or a placebo, and this would be given to them at the date of recruitment which will be from 0 to 12 months. Follow-up will be assessed every three months with the end of the trial at three years from the start of the study. Approximately 10% right-censored observations were simulated in each replicate dataset. This censoring criterion is achieved through a combination of specifying a scale parameter of 50 for the censoring distribution, a scale of 18 for the patient event times, keeping the majority of times within the interval from start to the end of study at 36 months.

Figure 3.8 is a histogram that represents the right-censoring approximation under the study design mentioned above. It shows that the censoring times are constant over time whereas the survival times are largely occurring before the end of the study. The overlapping area shows the frequency at which random censoring could occur if the censoring time is less than the survival time for a given patient. For this study, it was of interest to investigate the minimum sample size required to achieve a power of at least 90% for SNPs with a moderate effect of 0.4 and an EAF of 0.1. The treatment and interaction effect size used throughout will be 0.3 and 0.2, respectively. Two sets of analysis have been run. First, a SNP only analysis after which adjustment was made for

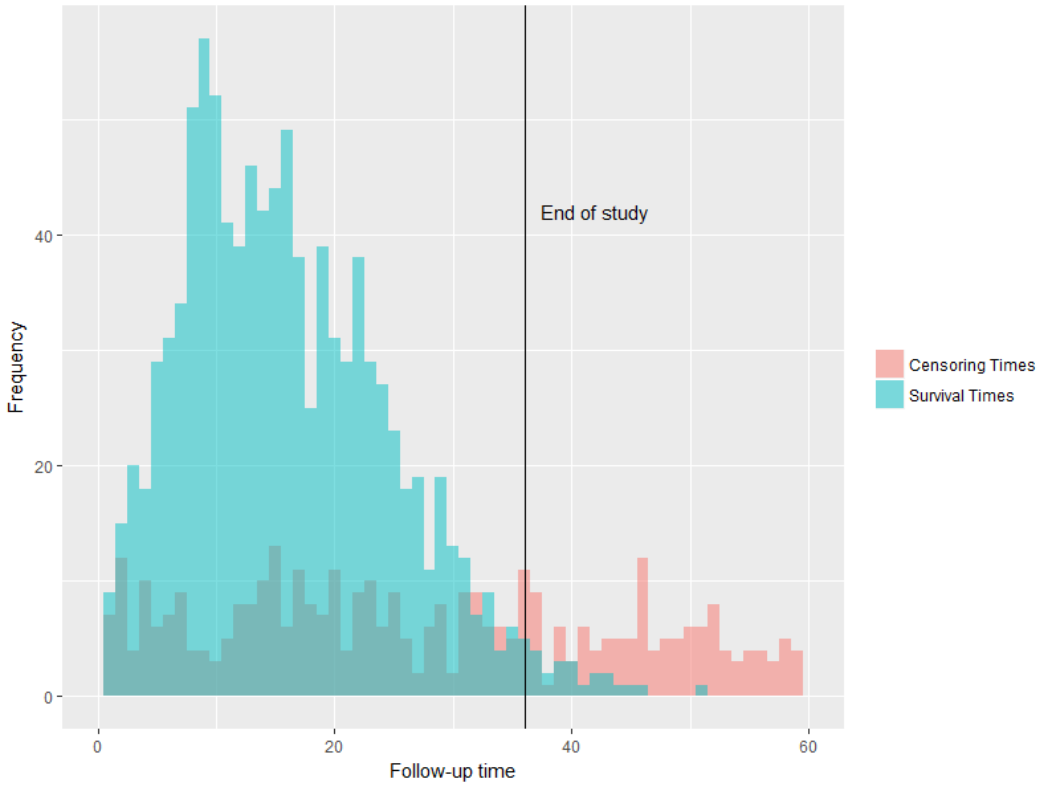


Figure 3.8: Histogram showing randomly simulated Weibull distribution estimates of survival and censoring times.

treatment and interaction for the second analysis. The second analysis featured power calculations for both the SNP effect and the joint association (Eq. 3.1). For all scenarios, assessment has been undertaken for the power to detect association using the Cox PH and Weibull regression models at genome-wide significance threshold $\left(5 \times 10^{-8}\right)$. For each scenario, 1000 simulations were performed by SurvivalGWAS_Power.

This example is not a complete representation of a study design protocol, but a demonstration of the use of the software. Investigators could use the tool for a more detailed comparison of methods and calculate sample size under different contributing factors. Different EAF thresholds, different percentage of censored observations, and varying effect sizes could be explored and how individually and collectively they will affect the power. These are all aspects to consider in a complete power calculation protocol. Figure 3.9 presents the input parameter tab of the software. The starting sample size is set at 400 for all calculations before adjusting and determining the desired sample size for a power of at least 90%. The example demonstrates the use of the Cox PH model

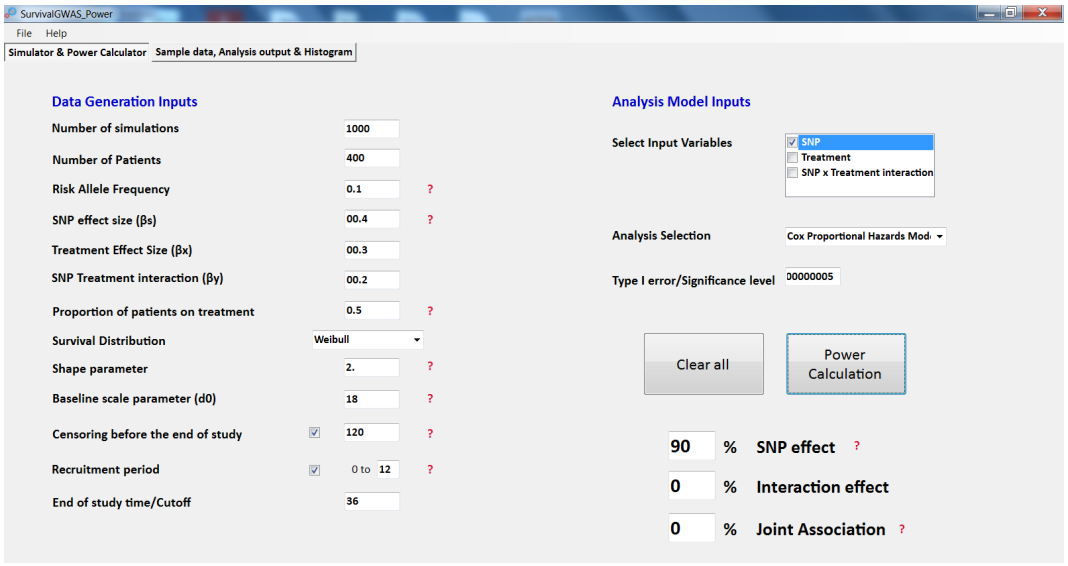


Figure 3.9: SurvivalGWAS_Power Example 1 input parameters. The example depicts a scenario with a sample size of 400, random right censoring and a recruitment period. Only a SNP effect is analysed within a Cox PH model.

considering only the SNP main effect. However the SNP, and treatment main effects, and a SNP-treatment interaction effect is taken into consideration in simulating event times. The analysis model is used to test the SNP association, i.e. the null hypothesis $H_0 : \phi_G = 0$ against the alternative $H_A : \phi_G \neq 0$, for which power is estimated to be 90% at a genome-wide significance threshold of $p < 5 \times 10^{-8}$.

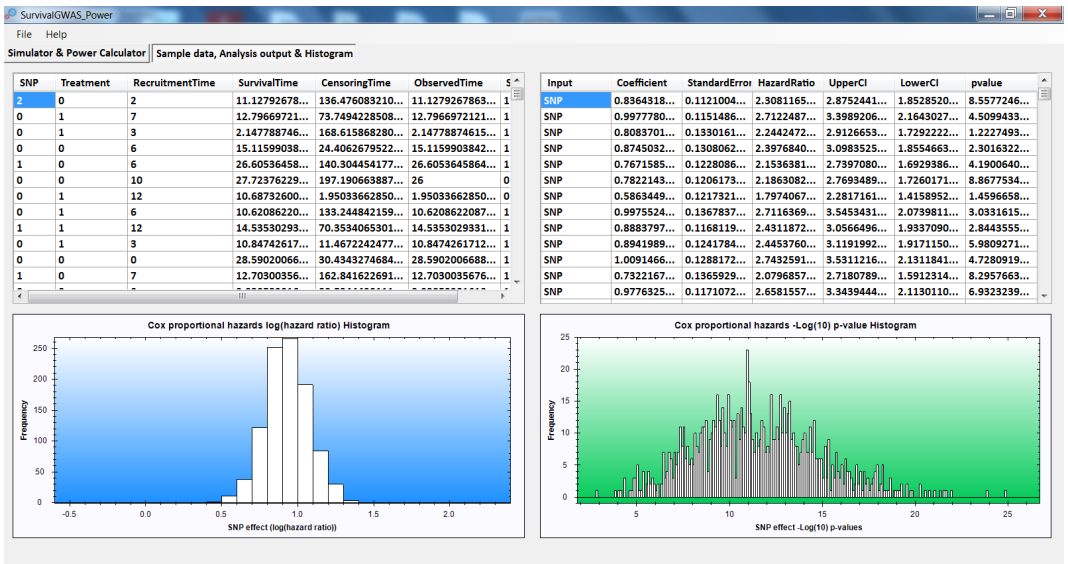


Figure 3.10: SurvivalGWAS_Power Example 1 power analysis output from Cox PH model. (Top left) Simulated sample dataset, (Top right) Parameter estimates of the SNP effect from each simulation run, (Bottom left) Histogram of SNP coefficient beta effects ($log(HR)$) across simulations & (Bottom right) Histogram of $-log_{10}$ p–values for the SNP effect across simulations.

Figure 3.10 shows the additional output from the analysis setup shown in Figure 3.9. The top left table displays one of the simulated datasets. Each row represents an individual patient. The dataset is a good way of checking that the censoring indicator and event times have been calculated correctly for the input parameters. The bottom left histogram shows the distribution of estimated SNP effect values (log(hazard ratio)) across simulations, which in this example are centred around 0.9, and not the true effect size of 0.8. This bias occurs as the data are simulated with a SNP-treatment interaction effect, that the analysis model does not take into account. The reason for the true log(HR) to be 0.8 and not the input parameter of 0.4 is that the shape of 2 used for simulating the event times increases the hazard multiplicative by 2.

The top right table shows the Cox PH analysis output from each simulation run, focussing only on the SNP effect. The bottom right histogram shows the $-log_{10}$ p-value for the SNP effect across simulations. Power, at the specified significance threshold, $\alpha < 5 \times 10^{-8}$, is approximated by the proportion of the 1000 simulations for which the p-value, $p < 5 \times 10^{-8}$ for the SNP effect on the outcome.

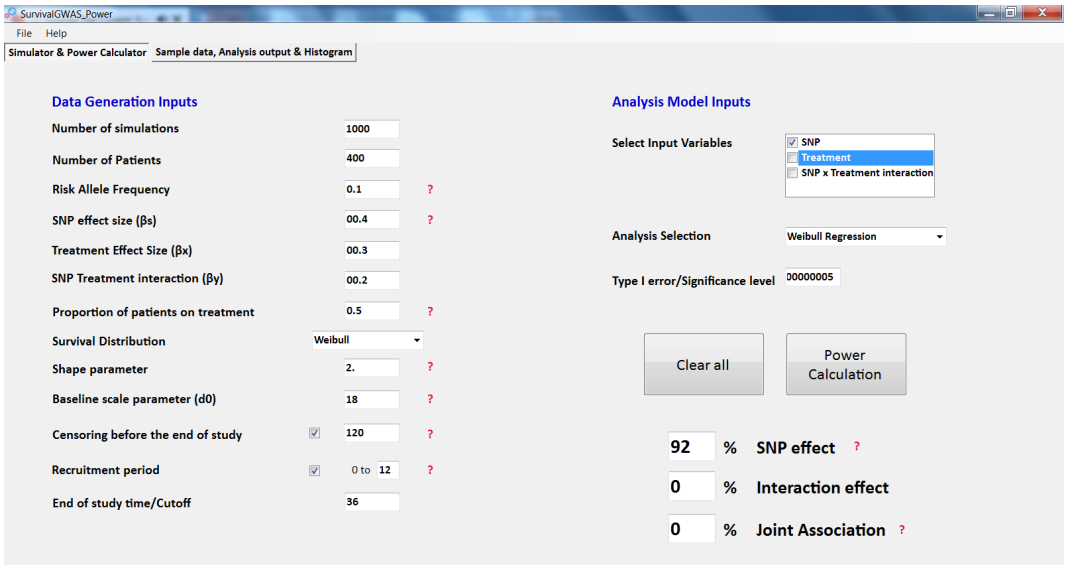


Figure 3.11: SurvivalGWAS_Power Example 2 input parameters. The example depicts a scenario with a sample size of 400, random right censoring and a recruitment period. Only a SNP effect is analysed within a Weibull regression model.

Figure 3.11 shows a power calculation based on the same input parameters shown in Figure 3.9. However, now assessing evidence of association using a Weibull regression

model. This model has a power of 92% which is greater than that of the Cox PH model. Figure 3.12 shows the output tab from the Weibull regression analysis adjusting

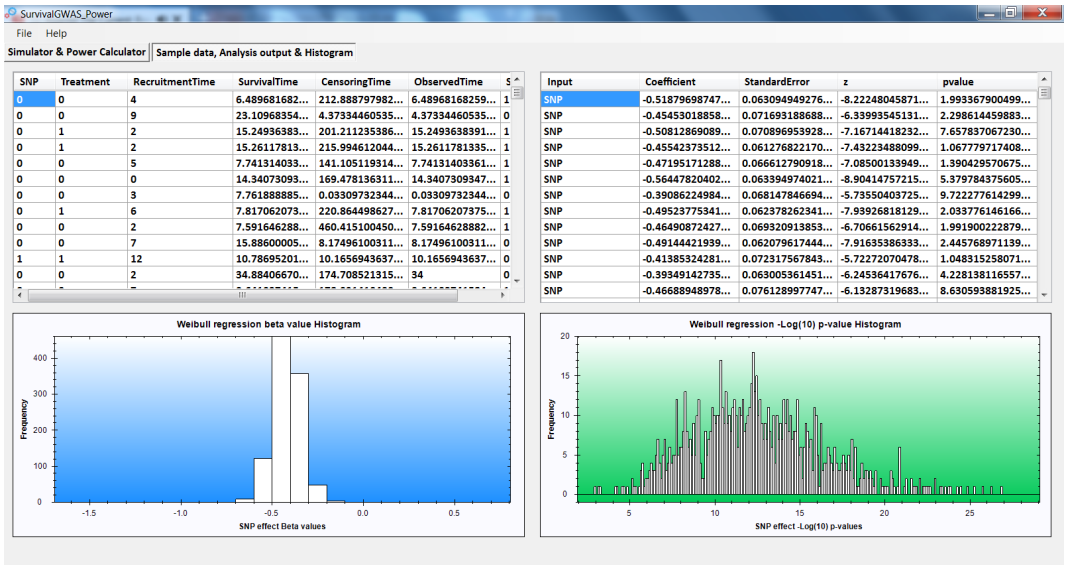


Figure 3.12: SurvivalGWAS_Power Example 2 power analysis output from Weibull regression model. (Top left) Simulated sample dataset, (Top right) Parameter estimates of the SNP effect from each simulation run, (Bottom left) Histogram of SNP coefficient beta effects ($log(AF)$) across simulations & (Bottom right) Histogram of $-log_{10}$ p–values for the SNP effect across simulations.

for only the SNP effect. The bottom left histogram shows the estimated SNP effect values (log(change in time)) across simulations. Converting the mean value of -0.45 to a hazard ratio the result is, $e^{0.45\times 2} = 2.459603, log(2.459603) = 0.9$. From this calculation, these estimates are also biased. To understand the bias generated by the incorrect models in contrast to fitting the correct model, Figure 3.13 presents the input parameter tab of the software adjusting for the SNP main effect along with treatment and an interaction effect. The same simulation model as in the previous example is used, including censoring during the study period and at the end of the study with a recruitment period. Event times are simulated with SNP and treatment main effects and a SNP-treatment interaction effect. From the analysis a Cox PH model is implemented to test: (i) the null hypothesis $H_0 : \phi_G = 0$ against the alternative $H_A : \phi_G \neq 0$, for which power is estimated to be 17% at a significance threshold of $p < 5\times 10^{-8}$; and (ii) the null hypothesis $H_0 : \phi_\gamma = 0$ against the alternative $H_A : \phi_\gamma \neq 0$, for which power is estimated to be 0% at a significance threshold of $p < 5 \times 10^{-8}$. However, what is most

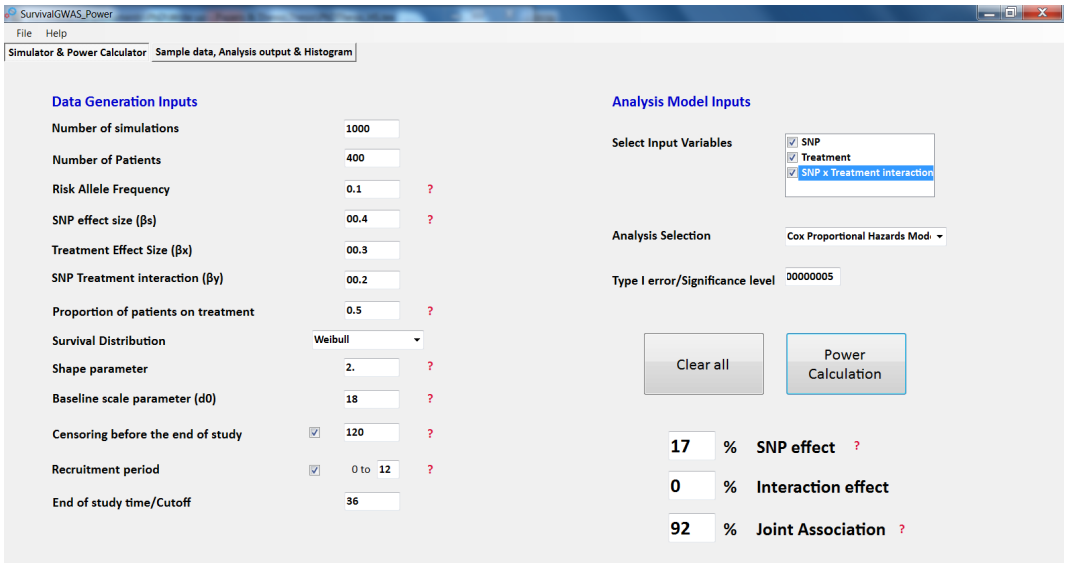


Figure 3.13: SurvivalGWAS_Power Example 3 input parameters. The example depicts a scenario with a sample size of 400, random right censoring and a recruitment period. A SNP, treatment and interaction effects, are analysed within a Cox PH model.

informative is the joint association test comparing the fit of two models. 92% of all tests indicate that the model fit with SNP, treatment and interaction terms is statistically significant at genome-wide significance.

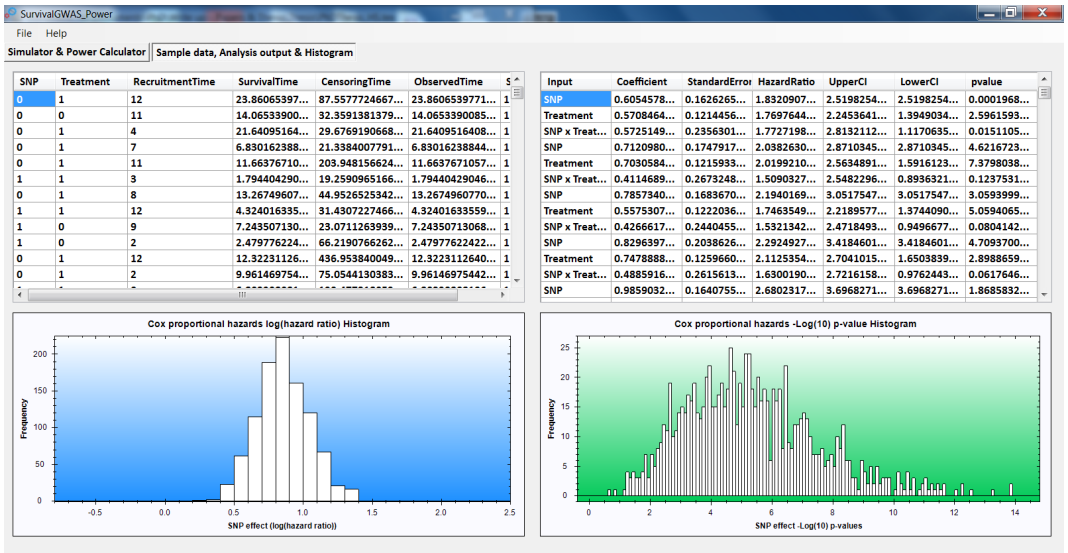


Figure 3.14: SurvivalGWAS_Power Example 3 power analysis output from Cox PH model. (Top left) Simulated sample dataset, (Top right) Parameter estimates of the SNP, treatment and interaction effects from each simulation run, (Bottom left) Histogram of SNP coefficient beta effects ($log(HR)$) across simulations & (Bottom right) Histogram of $-log_{10}$ p–values for the SNP effect across simulations.

Figure 3.14 shows the additional output from the analysis. This output corresponds

to the setup shown in Figure 3.13. The analysis output table in the top right shows the SNP, treatment and interaction output. The left histogram shows the distribution of estimated SNP effect values across simulations, which in this example are centred around 0.8. Unlike the power calculation based on analysis adjusting for only the SNP effect, the effect estimates are less biased.

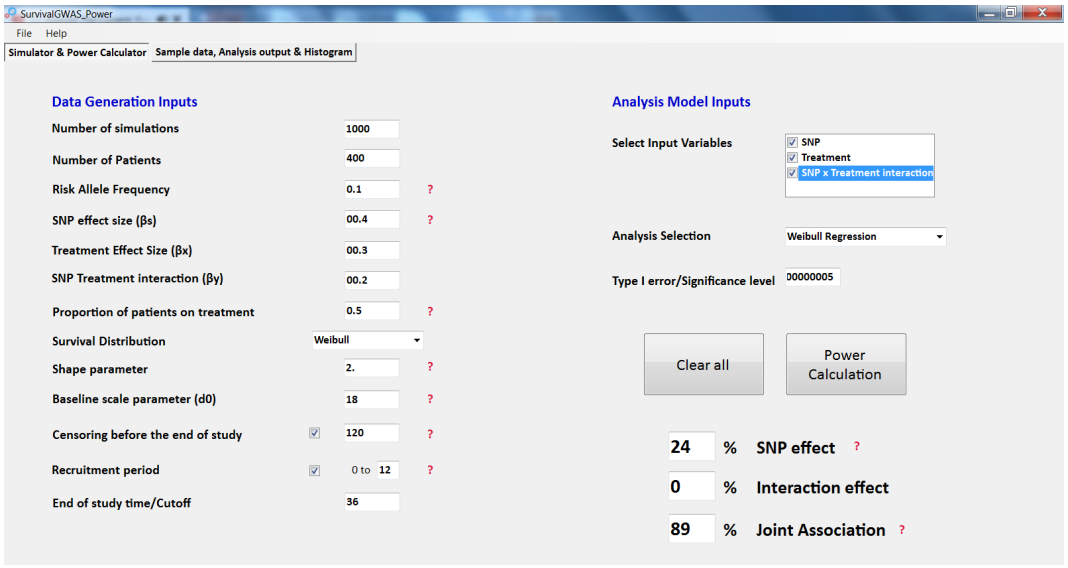


Figure 3.15: SurvivalGWAS_Power Example 4 input parameters. The example depicts a scenario with a sample size of 400, random right censoring and a recruitment period. A SNP, treatment and interaction effects, are analysed within a Weibull regression model.

A Weibull regression model has demonstrated to be the correct choice for analysis, as the regression-adjusted estimates are essentially unbiased showing a more precise estimation of the SNP effect size over the 1000 simulations (see Figure 3.16). The standard errors of the Weibull regression model with and without the inclusion of covariates are much smaller than the estimates from the Cox PH model. The addition of the treatment and interaction as covariates were minimally prognostic of our outcome, including them in the regression models explained some noise in the data.

The Weibull regression model adjusting for SNP, treatment and SNP-treatment interaction effects in the model obtained a power of 24% for the SNP effect compared to the 17% power estimated by the Cox PH model. Overall, an increase in sample size from 400 to 850 for this study is needed to achieve at least 90% power to detect associations

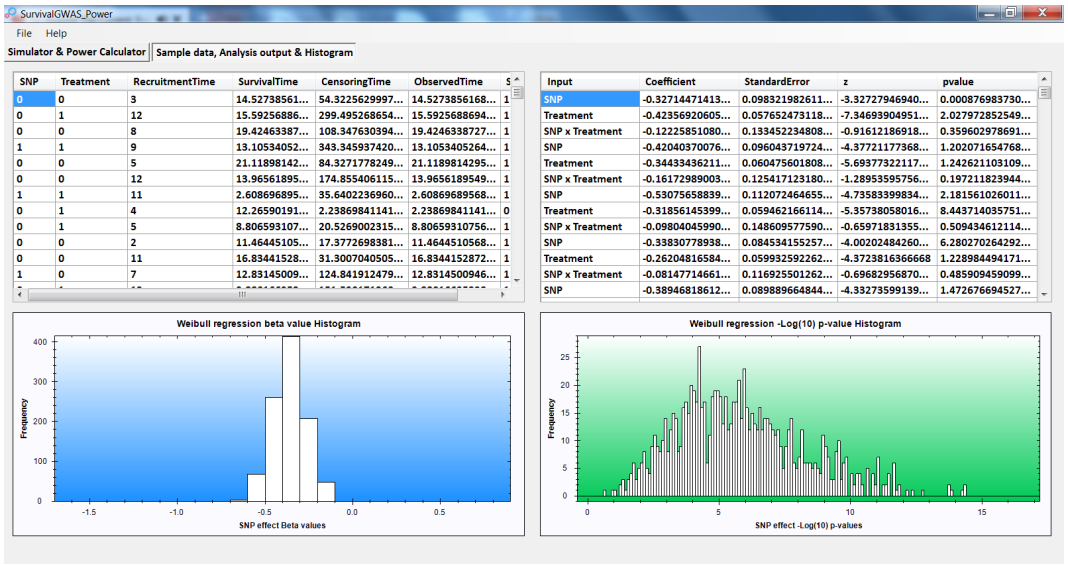


Figure 3.16: SurvivalGWAS_Power Example 4 power analysis output from Weibull regression model. (Top left) Simulated sample dataset, (Top right) Parameter estimates of the SNP, treatment and interaction effects from each simulation run, (Bottom left) Histogram of SNP coefficient beta effects ($log(AF)$) across simulations & (Bottom right) Histogram of $-log_{10}$ p–values for the SNP effect across simulations.

for the SNP effect at genome-wide significance using both the Weibull regression and Cox PH models adjusting for SNP, treatment and interaction effects.

As demonstrated by this example, SurvivalGWAS_Power can efficiently estimate the power of the Cox PH model and Weibull regression model under a variety of pharmacogenetic settings. Explicitly, allowing for testing of SNP main effects (i.e. testing the null hypothesis $H_0 : \phi_G = 0$ against the alternative $H_A : \phi_G \neq 0$), SNP-treatment interaction effects (i.e. testing the null hypothesis $H_0 : \phi_\gamma = 0$ against the alternative $H_A : \phi_\gamma \neq 0$) and LRT of the joint association model.

3.7 Discussion

In response to the lack of power calculation tools and the analytical bottleneck for identifying genetic factors associated with TTE data, the user-friendly tool, SurvivalGWAS_Power was developed. This program is the first to implement both data generation and power calculations for GWAS of TTE outcomes. The software is of particular relevance to pharmacogenetic studies, where the design will likely include alternative

treatment interventions and SNP-treatment interaction effects. However, the software is not exclusive to pharmacogenetic designs: for example, the treatment covariate can be used to represent any binary covariate and the event could be age of disease onset. This adds flexibility to the software for application to general GWAS of TTE outcomes.

SurvivalGWAS_Power can generate sample pharmacogenetic data with TTE outcomes over a range of study designs and perform power calculations using different analytical models. SurvivalGWAS_Power calculates the power to detect association of a SNP with a TTE outcome at a pre-specified significance threshold. The data can be analysed using a Cox PH model or Weibull regression model to account for non-proportional hazards.

To allow for flexibility of analysis using methods that are not currently supported by the power calculator, individual simulated data sets can also be output from the software. These datasets allow for users to simulate data and use other programs such as R for analysis. For example, Uno et al. (2014) have demonstrated that, where the PH assumption is invalid, the use of the Cox PH model will produce a loss of power to detect associations. They propose using robust alternative measures for the difference between survival curves instead of parametric models. The flexibility of our software enables generation of TTE data under models with non-PH that can be exported for association testing with methods supported by other software packages.

The project homepage can be found at the University of Liverpool, Statistical Genetics and Pharmacogenetics Research Group website (https://www.liverpool.ac.uk/translational-medicine/research/statistical-genetics/software/). The software is limited to Windows O/S users and is licensed under the GNU General Public License, version 3 (GPL-3.0). Therefore academics can edit the software to fit their requirements or use the code as a reference to help build similar tools.

This chapter explored the area of power and sample size calculation, introducing different analytical methods and original software. Once a study has been designed, the next step after the collection of data is the analysis procedure. Chapter 2 had already

expressed the need for analysis software for GWAS of TTE outcomes. The past, present and future perspective of this topic is explored in the next chapter.

CHAPTER 4

SINGLE VARIANT ANALYSIS OF GWAS WITH TIME-TO-EVENT OUTCOMES

4.1 Overview

Genome-wide association studies (GWAS) have revolutionised our understanding of the genetic basis of a wide variety of complex human traits and diseases. The focus of most GWAS has been towards binary phenotypes or quantitative traits, for which proficient software tools for analysis have been developed, such as SNPTEST (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html), PLINK (Purcell et al. 2007) and BOLT-LMM (Loh et al. 2015) (linear mixed models). However, for time-to-event (TTE) outcomes, very few computational analysis tools are available.

The main challenge, which explains why there currently is a lack of such powerful tools for survival analysis of GWAS is that the software is required to be computationally efficient while handling the scale and complexity of genetic data. The design of algorithms to achieve this requires advanced knowledge of programming pipelines and thorough examinations of current statistical genetics computational tools. Software should also offer use of a range of analytical models. This observation arose from the literature review in Section 2.2 whereby investigators would be inclined to using the Cox PH model for their study without indicating whether model checking was undertaken.

4.1.1 Objectives

This chapter seeks to evaluate methodology and software used for GWAS with TTE outcomes. This review is key to understanding more about the current availability of computational tools for GWAS of TTE outcomes. This chapter briefly describes popular

software for binary and quantitative traits, which inevitably helps in the design of a suitable computational analysis tool which draws on the strengths of other software. The ultimate aim of this chapter is to develop and test through simulations a novel computational analysis tool for GWAS with TTE outcomes.

4.2 A Review of Genome-Wide Time-to-Event Studies

In Section 2.2, a comprehensive literature review of pharmacogenetic studies was conducted, covering both genome-wide and candidate gene studies. However, expanding on the literature review, by analysing papers outside this context, we can gain a broader understanding of the different types of TTE phenotypes analysed, study designs reflecting different censoring options, the underlying genetic models used and imputation software. As a result, this information provides us with a better understanding of the specifications for which our analysis software should be built. There are many recent GWAS published with the focus on survival outcomes. Table 4.1 summarises key findings from a selection of recent studies.

Taking into account both the information in Tables 2.1 and Table 4.1, it can be observed that: (i) right censoring is the most common type of censoring; (ii) the additive genetic model on the log hazard ratio is assumed more often than the recessive and dominant models; and (iii) the Cox PH model is applied in all studies. The biggest discrepancy between each study is the choice of imputation software. This disparity is very important as different imputation programs produce different output file formats, and therefore association analysis software is required to handle a wide range of genotype file formats. Computational tools need to read in and transform genotype probability data using genetic dosage models accounting for genotype uncertainty.

Examining the latest version of each software, IMPUTE2 (Howie et al. 2012) outputs genotype files in GEN (.gen) format, Beagle 4.1 (Browning & Browning 2016) and minimac3 (Das et al. 2016) both output variant call format (VCF) files. Each program uses its own input data format and algorithms to perform the imputation of

Study	Sample Size	Variants	Censoring	Phenotype	Method	Imputation Software	Analysis Software
Walter et al. (2011)	16,995	2.5 million SNPs	Right	(i) All-cause mortality. (ii) Survival free of major disease or death.	CPHM. Additive genetic model.	Unknown	Unknown
Johnson et al. (2016)	3256	≈4 million SNPs in each of 4 cohorts after QC.	Left, Right	Overall survival.	Multivariate CPHM.	IMPUTE2	R 3.1.3
Phipps et al. (2016)	3494	≈2.7 million SNPs	Right	Overall survival. CRC specific survival.	CPHM. Log additive genetic model.	MaCH	R 2.15.3
Kapoor et al. (2014)	1788	4,058,415 SNPs	Right	Age at onset	CPHM. Log additive genetic model.	Beagle 3.3.1	R - '*Survival*' package
Wu et al. (2014)	1005	5,038,636 SNPs	Right	Overall survival	CPHM. Additive genetic model.	MaCH	SAS
He et al. (2016)	1962	5,918,992 SNPs	Right	Time to; (i) smoking initiation, (ii) persistent smoking, (iii) tolerance, (iv) cessation.	CPHM	IMPUTE2	R - '*coxme*' package.

Table 4.1: Summary of GWAS with TTE outcomes. Abbreviations: CPHM, Cox proportional hazards model; CRC, colorectal cancer; QC, quality control; SNPs, single nucleotide polymorphisms.

non-genotyped single nucleotide polymorphisms (SNPs). It is possible to convert between file types using custom-built software such as PLINK 1.9 (Chang et al. 2015), GTOOL (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html), BCFtools (Li 2011) or DosageConvertor (https://genome.sph.umich.edu/wiki/DosageConvertor).

Analysis software should offer compatibility with many of the different file types produced from imputation. GEN and VCF files are the most frequently used for analysis before and after imputation. All of the file types that contain genotype information can be compressed to save space, most commonly gzipped (.gz) or bgzipped (.bgz). Computational tools that are able to read the compressed files directly save the user time and computer storage space.

One final observation is that, in all the studies, adjustment was made for multiple covariates to account for potential confounding variables. Pairing this with the analysis of millions of SNPs and thousands of individuals, the computational burden is increased. Analytical software needs to be able to perform millions of individual tests, whilst maintaining computational efficiency.

4.3 Extensions of Time-to-Event Models for GWAS

Statistical methodology for GWAS of TTE outcomes is a rapidly developing area of research, and there is not as yet a consensus as to the most effective methods. Traditional methodologies within genetic research, indicated earlier in Chapter 2, has been limited to use of the Cox PH model and log-rank test. Alone these models are unable to deal with the complexities associated with the modelling of relationships between genetic biomarkers and TTE outcomes. Outside these traditional analyses, the methodology has been developed to cater for a wider range of TTE outcomes.

Subirana & González (2013) highlighted the importance of accounting for genotype imputation uncertainty within survival models. They compare three approaches through a simulation study; a naive (directly typed best guess), dosage and latent class (maximisation of SNP probabilities in likelihood) approach. The first two are implemented

in a Cox PH model and the third within a Weibull regression model. Performance of both the dosage model and latent class approach are very similar when comparing bias, mean squared error, power to detect associations and coverage.

Lin et al. (2011) had evaluated the single SNP approach against a kernel machine SNP-set Cox PH analysis, concluding that the latter is a more robust choice, in that it suffers a little loss in power when the effect of the SNP is linear but is useful when the effects of the SNPs are more complex or when epistasis[1] is present. This method is most helpful when multiple SNPs are causal within a region, indicating a greater application to the multi-SNP analysis of common variants.

Vandin et al. (2015) discussed using standard methods, with the implementation of the log-rank test for GWAS. There are obvious problems with the use of the log-rank test such as its failure to handle unbalanced populations[2] resulting in many false positive/negative associations and its inability to adjust for covariates. The paper continues by providing an alternative algorithm called "ExaLT" which computes a *p*-value under an exact permutational distribution. This algorithm, however, is unlikely to be sufficiently computationally efficient for GWAS. It is also unclear how the proposed algorithm accounts for continuous confounders.

ExaLT was proven to be beneficial over the standard log-rank test for genomic data, and therefore should be considered as an option to testing the association between two groups. Even with this method, the Cox PH model is still considered the optimal choice for analysis when considering adjustment of covariates and computational efficiency. The methods described by Subirana & González (2013), Lin et al. (2011), Vandin et al. (2015) have been considered alongside frequently used survival models for implementation into software.

[1]Epistasis is the interaction between multiple genes.

[2]In a genomics study, the two groups are defined by a SNP, such as comparing carriers and non-carriers of an allele. The sizes of the groups in many of these studies are unbalanced; one group is usually much larger than the other.

4.4 Time-to-Event Analysis Tools in Genetic Research

Many of the recent GWAS published (displayed in Table 4.1) with a focus on survival outcomes conducted analyses using standard statistical software, such as R or SAS. These programs are limited as they need a lot of available random access memory (RAM) and time to load large data files. A standard four core laptop would take weeks to run an analysis of one million SNPs and over a thousand samples, with the possibility of the computer running out of memory. To help avoid this issue high-performance computing (HPC) clusters are used to improve efficiency and provide an increased data storage capacity, however, R and SAS are not easily amenable to these solutions or capable of handling large-scale GWAS data.

Programs such as ProbABEL (Aulchenko et al. 2010) were explicitly created to tackle this problem, though users have flagged many difficulties. ProbABEL is described as a software package for the analysis of genome-wide imputed SNP data and quantitative, binary, and TTE outcomes. Nevertheless, it is limited to the use of only the Cox PH model for TTE data. The output parameter of particular importance in all types of studies is the p-value, which ProbABEL does not output. The software does output the coefficient estimates, standard errors and log-likelihood. This implies the user is required to calculate the p-values themselves, using a Wald or likelihood ratio test (LRT). However, ProbABEL is currently maintained, and therefore improvements in future releases may correct many of these problems.

Genipe, created by Lemieux Perreault et al. (2016), is a new pipeline for imputation with automatic reporting of output from various statistical analyses. Genipe implements existing approaches to imputation and then utilises the python package '*lifelines*' to generate association summary statistics via the Cox PH model. This tool is useful but is restricted because it relies on existing software such as PLINK and IMPUTE2, and is limited to the Cox PH model.

As previously mentioned in Section 1.4, the success of a computational analysis tool

relies on the choice of programming language and environment. State of the art software (post-2013) such as SNPTEST, PLINK 1.9, BOLT-LMM (Loh et al. 2015), EPACTS (https://genome.sph.umich.edu/wiki/EPACTS) are all developed using C++. Many other GWAS software such as PyLMM (http://genetics.cs.ucla.edu/pylmm/) are written using Python. Python is considered to be slower than the C based languages, but ultimately speed is relative to the design of the algorithms and cleanliness of written code. These languages are chosen due to the convenience of running on the Linux operating system (O/S) and HPC clusters. In Chapter 3, SurvivalGWAS_Power was created using C# due to the flexibility it provided for using .NET libraries. C# is considered faster than Python but slower than C++ when compiling code on Linux machines. C# is hindered further due to the use of Mono, a third party compiler. This decrease in performance may change with the introduction of Visual Studio Code and .NET Core (https://code.visualstudio.com/docs/other/dotnet), which provides users with a "blazing fast and modular platform for creating server applications that run on Windows, Linux and Mac".

Due to the inadequacy of current GWAS analysis tools for TTE outcomes, the software tool SurvivalGWAS_SV has been developed, which has addressed the difficulties that are faced by other programs and currently employs a single SNP analysis approach using two commonly used survival analysis models.

4.5 SurvivalGWAS_SV

4.5.1 Implementation

SurvivalGWAS_SV is a freely available program created for the analysis of GWAS of imputed genotypes with TTE outcomes. It is the second program to be released under the SurvivalGWAS suite of software, which also includes the complementary power calculator "SurvivalGWAS_Power" described in detail in Chapter 3. SurvivalGWAS_-SV is a C# developed program; the executable file (.exe) with all dependencies can be quickly distributed for any O/S. When used on Linux machines it is necessary to run

the executable using the compiler Mono (http://www.mono-project.com/download/) or the untested .NET core.

Key features include: (i) compatibility with the file formats produced by programs such as IMPUTE2, thereby directly accommodating imputed data without the need for file conversion; (ii) a range of survival analysis models are available with the foundation in place for implementing additional methods as required; (iii) options for testing SNP-covariate interactions, presenting p-values for the entire model and individual covariate tests of association; and (iv) compatibility with HPC clusters.

SurvivalGWAS_SV has undergone many changes over the last two years. The first version deployed had a sequential process pipeline, where each SNP was processed one at a time, analysed and with output generated. Figure 4.1 depicts the multithreaded analysis process implemented in the most recent version of SurvivalGWAS_SV. Improvements to the software over the years have been to cater for the continual change in data file types, reflecting the addition of VCF files. Updating the software to reflect the changes within the field should increase the number of users of the software.

The multithreading pattern uses parallel tasks and a concurrent queuing system (reader-writer lock[3]) in places to process a sequence of input values. The threads are completely independent, to avoid shared resources between threads, which can result in overwriting parameters. Each thread executes a validation and analysis protocol for a given batch, and the queue for writing to the output file acts as a 'buffer' that allows only one thread to write at a time otherwise threads will be writing on the same line. Whichever thread finishes first will write to the output file and then move on to the next line of the input file. A practical example for visualising this multi-threading pipeline would be to consider multiple assembly lines in a factory. Each item in the assembly line is taking resources from the same location. Each line produces a fully assembled product. However, there is only one truck that distributes the products. The product placement order in the truck depends on whichever assembly line finishes first.

[3]A reader-writer lock within a program allows a single thread access to a file that multiple threads have access to. This system is in place to avoid overwriting of file contents.

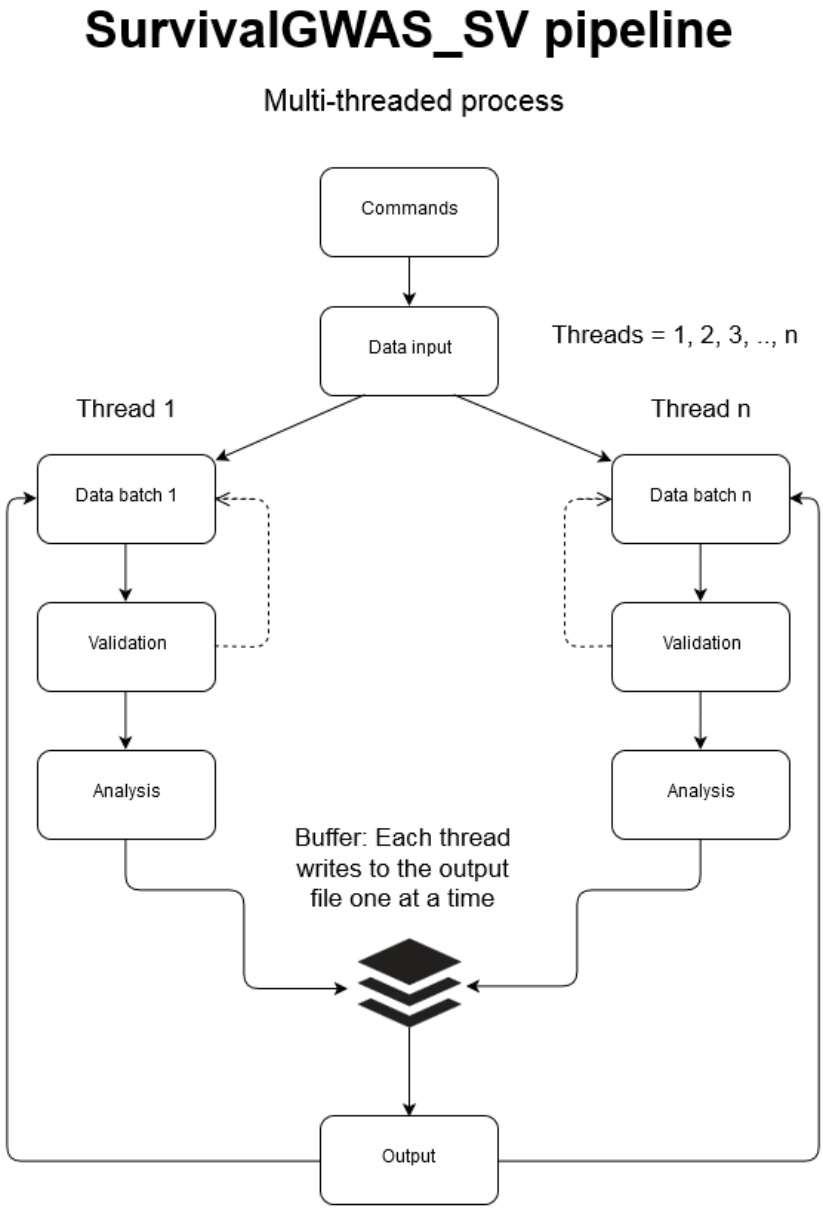


Figure 4.1: Flowchart of SurvivalGWAS_SV analysis process.

4.5.2 User Interface

SurvivalGWAS_SV is a console application utilising command line inputs. The software is run from a command prompt terminal, compatible with Linux, Windows and Mac OSX. The program requires little interaction from the user since a script of commands can be submitted to the program, which is useful for the analysis of large data files. The user can specify "batches" of the data file to analyse in parallel using multiple compute nodes, where each core can run a different part of the analysis. The program re-

quires Mono to run the software on Linux and Mac OSX. Figure 4.2 shows the execution of the software through MobaXterm; a windows desktop terminal for the remote server and SSH client access. The screen-shot shows the printed description of the software in the terminal after executing the software command; mono survivalgwas-sv.exe.

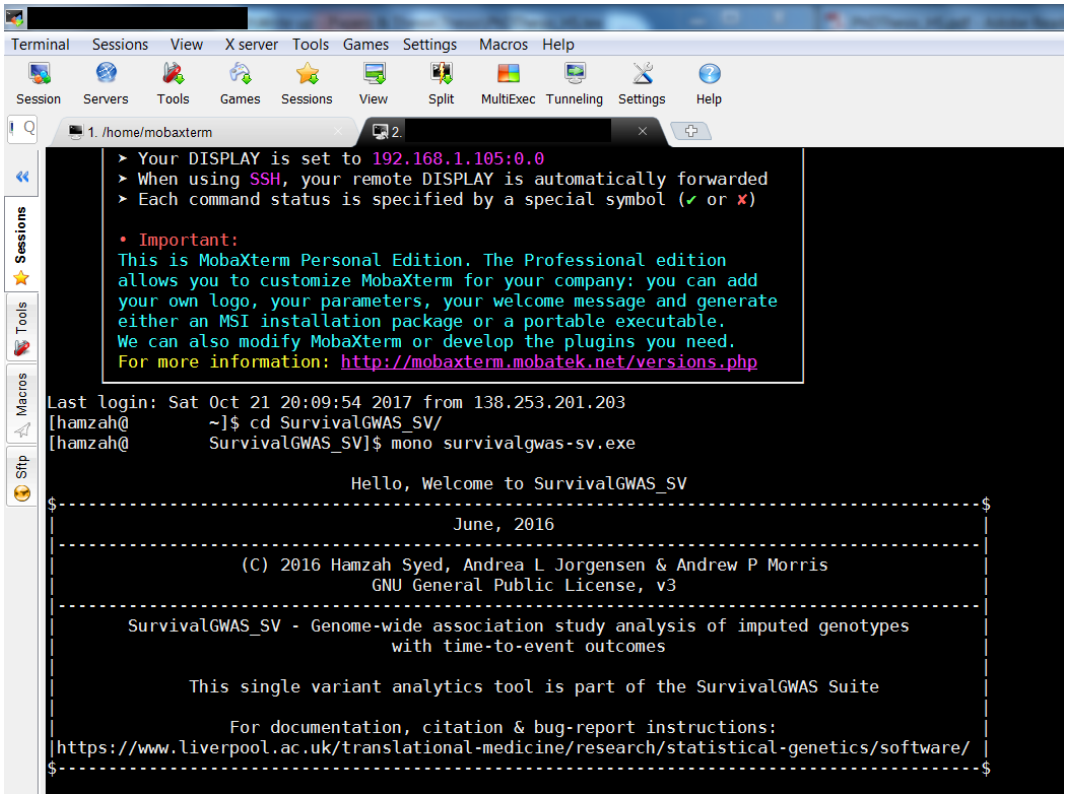


Figure 4.2: Using SurvivalGWAS_SV through MobaXterm. Change directory to where the SurvivalGWAS_SV executable and libraries are saved. Run executable using Mono.

4.5.3 Inputs

SurvivalGWAS_SV is designed in a very simplistic way. First, the user is required to specify the two data files that will be read into the program. The first file must be a genotype file (.gen or .impute) or a VCF text file that contains SNP genotype probabilities (imputed or non-imputed), and the second file should be a sample file (.sample) that contains all the covariate, survival time and censoring indicator information for each individual. The software supports VCF files containing the SNP genotype probabilities, dosages and/or hard genotype calls. In some circumstances, the user would have the genotype files compressed, either in a .zip or .gz file format, both of which can be read

into the software directly. Second, the user specifies details about terms to include in their analysis model, such as covariates and/or interactions, whilst also specifying the censoring indicator and observed survival time. Third, the user must specify the range of SNPs to be analysed, to enable efficient parallel computing in batches. Last, the user must enter the chosen analytical method to use and the name of the file to which the analysis output will be saved. If the user is analysing covariates within the model, but does not require summary statistics for the covariates to be included in the output file, an option is available for only printing the results for the SNP or interaction effects. This option is helpful when creating graphical summaries, such as Manhattan plots, using other programs.

GEN file	
SNP ID	This entry is usually used to denote the chromosome number or 'SNP<number>'.
rs number	This entry is an ID number which uniquely identifies the genotyped SNP. A rs number is an accession number used by researchers and databases to refer to specific SNPs. It stands for 'Reference SNP cluster ID'.
Base pair position of the SNP	A value that describes the SNPs position on the chromosome.
Major allele	The entry will be A, C, T or G, i.e. corresponding to the four possible nucleotides.
Minor allele	The entry will be A, C, T or G.

Table 4.2: GEN file contents. Abbreviations: SNP, single nucleotide polymorphism.

Table 4.2 details the contents of the genotype file. The first column in the file is the SNP number, the second is the unique SNP identifier number and the third is the base-pair position of the SNP. Columns four and five are the two alleles. These alleles represent the major and minor alleles, however, often they are referred to as the reference (REF) and alternative (ALT) alleles in the GEN and VCF files. The remaining columns on each row represent the genotype of each individual at the SNP. Each individual will have three entries, representing their probability of having the

major homozygous, heterozygous and minor homozygous genotypes respectively. If there are imputed SNPs in the dataset then the probability for each genotype will be between 0 and 1, whereas if the data is not imputed then two of the possible genotypes will be 0 and the other will be a 1, strictly defining this genotype for an individual as 100% certain. SurvivalGWAS_SV assumes an additive genetic model (mode of inheritance) and therefore converts these probabilities into dosages for the purpose of fitting models to test for association. For example, in Script 4.1, each row provides information on two individuals. At SNP1 the two alleles are A and G therefore the set of 3 probabilities correspond to the genotypes AA, AG and GG respectively. The first individual has a probability of 0.9 for the genotype at SNP1 to be AA, 0.1 for AG and 0 for GG. This genotype uncertainty needs to accounted for in the analysis models. The second individual has a definitive probability of 1 for genotype AA.

```
SNP1 rs1 1000 A G 0.9 0.1 0 1 0 0
SNP2 rs2 2000 G T 0 0.65 0.35 0 0.99 0.01
SNP3 rs3 3000 C T 1 0 0 0 1 0
SNP4 rs4 4000 A C 0.04 0.8 0.16 1 0 0
SNP5 rs5 5000 A G 0 1 0 0 0 1
```

Script 4.1: Imputed GEN file contents for five SNPs and two individuals.

```
## fileformat=VCFv4.1
## filedate=2016.12.14
## source=Minimac3, PLINK, VCFtools
## contig=$<$ID=1$>$
## FORMAT=$<$ID=GT,Number=1,Type=String,Description="Genotype"$>$
## FORMAT=$<$ID=DS,Number=1,Type=Float,Description="Estimated
    Alternate Allele Dosage : [P(0/1)+2*P(1/1)]"$>$
## FORMAT=$<$ID=GP,Number=3,Type=Float,Description="Estimated
    Posterior Probabilities for Genotypes 0/0, 0/1 and 1/1"$>$
## INFO=$<$ID=AF,Number=1,Type=Float,Description="Estimated
    Alternate Allele Frequency"$>$
## INFO=$<$ID=MAF,Number=1,Type=Float,Description="Estimated Minor
    Allele Frequency"$>$
```

```
10 ## INFO=$<$ID=R2,Number=1,Type=Float,Description="Estimated
       Imputation Accuracy"$>$
11 # CHROM  POS  ID  REF  ALT  QUAL FILTER INFO FORMAT
12 1 1000 rs1 A G . . AF=0.15;MAF=0.15;R2=0.00682 GT:DS:GP
       0/0:0.100:0.900,0.100,0.000 0/0:0.000:1.000,0.000,0.000
```

Script 4.2: VCF file contents example for one variant and two individuals.

Sample file	
ID_1	This column is a patient identifier for each individual.
ID_2	This column provides an option to give a second patient identifier for each individual. You may wish to do this to identify study site, for a multi-site study.
missing	This column details the proportion of missing genotype data for the individual.
age	This column includes a covariate representing the individual's age. This covariate is quantitative.
gender	This column includes a covariate representing the individual's gender. The covariate is binary, coded 0 (female) and 1 (male).
bmi	This column includes a covariate representing the individual's Body mass index (BMI). This covariate is quantitative.
treatment	This column includes a covariate representing which treatment the individual has been prescribed. The covariate is binary, coded 1 (treatment A) and 0 (treatment B).
time_to_event	This column contains the observed survival time for each patient after right censoring.
status	This column includes the outcome for the individual. In this case, our outcome is 'event' represented by 1 if the event of has occurred or 0 if the event has not occurred/censored.

Table 4.3: Example contents of a sample file. Not all sample files contain these exact columns.

With the use of next generation sequencing (NGS) technology, the new standard for storage of genotypes is with VCF files. The same example as above for SNP1 is represented in Script 4.2. The first few lines from the top of the sample VCF file starting with a ##, are meta-information lines, as shown in Script 4.2. These lines are subsequently followed by the header line then the genotype information for each sample. The meta-information provides details on the the VCF version number, the source

program which created the file amongst other information unique to each file. The FORMAT and INFO rows describe the variables used within the genotype information. This information includes MAF or if genotype probabilities are included in the file. The header groups all this information and the genotype information of a variant for each sample in a single row. A detailed explanation on the latest VCF file versions and their contents is provided on the Samtools github page (https://github.com/samtools/hts-specs).

A sample file includes one row per individual and can include as many covariates as required, each represented by a different column. Covariates can be binary or quantitative. Some individuals have missing covariate values which should be coded as "NA", for SurvivalGWAS_SV to distinguish which values are missing and which are not. These individuals will be excluded from analyses which adjust for that covariate. Table 4.3 provides a description of the possible columns within a sample file. The mandatory columns are the "time_to_event" and "status" columns.

4.5.4 Algorithms and Validation

Before the data can be analysed, a number of conversions and quality control measures must be performed by the software. When the genotype file is read in, one SNP at a time, either directly typed or imputed, SurvivalGWAS_SV will convert the genotype probabilities for each individual into a "dosage" under an additive model for the alternative allele. This model enables appropriate analysis for imputed SNP data by taking account of the uncertainty in the imputation process. The dosage model is given by Eq. 1.2. The implied ordering in the file for genotype probabilities is the reference homozygote, heterozygote, alternate homozygote. Note that the values are forced to sum to 1.00 exactly, and only bi-allelic variants are considered.

SurvivalGWAS_SV throws exemptions whenever the user has specified an incorrect command or states a variable name that cannot be found in the data files. In such an event, the program will exit the application and will require re-submission of the

task. The program also handles missing values within the sample file. If an individual has missing values (in the form of "NA") for survival time, censoring indicator or a covariate used in the model, then the individual is removed from the analysis with their corresponding SNP information.

VCF and GEN files are handled differently within SurvivalGWAS_SV because VCF files contain more information than GEN files. The files differ in a number of important ways: (i) genotype information within VCF files are given in hard genotype calls, dosages and/or probabilities; (ii) VCF files contain an estimation of MAF for each variant; and (iii) sample identification numbers are provided in a VCF file. A lot of this information needs to be filtered out or skipped over to access the data that is required.

4.5.5 Statistical Methodology

Analysis can be carried out using one of two methods: (i) a Cox PH model; or (ii) a parametric Weibull regression model. Both methods have their advantages under different scenarios. More details are discussed in Section 1.3. As discussed in Section 4.3, the contribution of Subirana & González (2013) provided a clear foundation for us to build upon, influencing the decision to implement the Cox PH and Weibull regression models, with an additive dosage model, into SurvivalGWAS_SV. Most importantly, the SNP-covariate interaction effects are modelled using the joint association test outlined in Eq. 3.1, replacing the treatment covariate with a vector of covariates, $\hat{\mathbf{x}}_\mathbf{i}$ and the SNP-treatment effect with a SNP-covariate interaction for the users' covariates of choice.

4.5.6 Usage Commands

```
$ mono survivalgwas-sv.exe -gf= -sf= -threads= -t= -c= -cov= -i= -
    chr= -lstart= -lstop= -method= -p= -o=
```

Script 4.3: SurvivalGWAS_SV command line example without defined parameters.

In the example shown in Script 4.3, each command is separated by a space and begins with '-' and ends with '=' before specifying the selected option. Table 4.4 outlines

the syntax for each command in SurvivalGWAS_SV with their corresponding usage description. The "-threads=" command is a recent introduction using the *'semaphore'* class in C#. This command substantially improves the speed of analysis. For example, a user has availability of a single computing node with 8 cores on a given cluster. The cluster is interactive, and therefore requires execution of the program on the command line. If the user wishes to analyse 10,000 SNPs, and they specify "-threads=5", the software will automatically assign computing resources analysing the 10,000 SNPs in 5 equal batches, across one or more computing cores. However if the user were to create more threads than is required, this can slow down the analysis, because threads will start to queue to use resources which are bound by reader-writer locks.

Command	**Description**
-gf=	This specifies the genotype file. Typically a .gen, .vcf, .gen.gz.
-sf=	This specifies the sample file (.sample).
-threads=	Specifies the number of threads. On a multi-core system, multiple, threads can execute tasks in parallel, with each core executing a different thread or multiple threads.
-t=	This specifies the time to event (column heading name) in the sample file.
-c=	This specifies the censoring indicator/outcome in the sample file.
-cov=	This specifies the covariates to adjust for in the model. Each one separated by a comma (,). e.g. -cov=cov1,cov2,cov3. Note: Categorical variables need to be converted to binary as software only assumes continuous or binary covariates.
-i=	This specifies the interaction between the SNP and a covariate. Separate using a comma (,). e.g. –i=SNP,Treatment
-lstart=	This specifies the line in the genotype file at which the start position of analysis will occur. Used to break large files into small batches for parallel computing.

Table 4.4 continued from previous page

-lstop=	This specifies the line in the genotype file at which the end position of analysis will occur. Typically the number of lines is equal to the number of SNPs in the file.
-sp=	The start position (in base pairs) on the chromosome. Still need to specify the number of lines in the file using -lstart & -lstop commands. This is an optional command.
-ep=	The stop position (in base pairs) on the chromosome. This is an optional command.
-chr=	This specifies the chromosome number to be output in the text file.
-p=	Enter "onlysnp" if only the results from the SNP analysis are to be output and "onlyint" if only the results from the SNP-covariate interaction analysis are to be output. Otherwise, the output will have separate rows for the SNP and all adjusted covariates.
-m=	This specifies the method for analysis. The choice is either "cox" for the Cox PH model or "weibull" for the parametric Weibull regression model.
-o=	This specifies the name of the file for output to be saved in. e.g. name.txt
-help	Outputs a full list of commands and usage help.

Table 4.4: List of commands available in SurvivalGWAS_SV and their corresponding usage description.

Assuming all data files and software are in the same folder, the command line in a Linux terminal for the analysis of 10000 SNPs and two additional covariates using a Cox PH model is shown in Script 4.4.

```
1 $ mono survivalgwas-sv.exe -gf=data.gen -sf=data.sample -t=event\
    _times -c=censoring -cov=covariate1,covariate2 -chr=1 -lstart=0 -
```

```
    lstop=10000 -m=cox -p=onlysnp -o=output.txt
```

Script 4.4: SurvivalGWAS_SV command line example with dummy input parameters.

The user can specify the exact location of the data files and where the output file will be saved. e.g. /DIRECTORY/DATA/output.txt. Script 4.5 shows an example of a shell script (.sh) based on the same parameters displayed in Script 4.4. The purpose of this shell script is to distribute the analyses between 10 computer cores within a Linux cluster (Sun-Grid Engine Batch System). These files are usually labelled "example.sh".

```
1  #!/bin/bash
2  #$ -o stdout
3  #$ -e stderr
4
5  DIRECTORY=/SurvivalGWAS\_SV #Location of software and data
6  str1=0 #Start position in genotype file
7  str=10000 #Number of SNPs/lines in genotype file
8  no_of_jobs=10 #Number of cores
9  inc=`expr \( $str - $str1 \) \/ $no_of_jobs` #Increment
10
11 #SGE_TASK_ID takes values 1:no_of_jobs
12 nstart=`expr \( $SGE_TASK_ID - 1 \) \* \$inc'
13 nstop=`expr $nstart + $inc - 1'
14 mono $DIRECTORY/survivalgwas-sv.exe -gf=$DIRECTORY/data.gen -sf=
       $DIRECTORY/data.sample
15 -t=event_times -c=censoring -cov=covariate1,covariate2 -chr=1 -
       lstart=$nstart -lstop=$nstop -m=cox -p=onlysnp
16 -o=$DIRECTORY/output${SGE_TASK_ID}.txt
```

Script 4.5: Shell script for running SurvivalGWAS_SV on a HPC cluster. Comments are highlighted in green.

```
1 $ qsub -t 1:10 example.sh
```

Script 4.6: Multitple core submission example using the UNIX 'qsub' command.

To submit the shell script the command line shown in Script 4.6 should be used. This command submits the analysis and divides the workload over ten cores. This distribution of data and analyses will produce ten output files which can be concatenated using the "cat" command in Linux (example provided in Script 4.7). Joining the files will also add multiple header lines to the file. Therefore it is important to delete the duplicate header lines before or after concatenation so that errors do not occur when producing a Manhattan plot in R.

```
$ cat file1 file2 file3 > jointfile.txt
```

Script 4.7: Concatenation of multiple files using the UNIX 'cat' command.

4.5.7 Output

The format in which the output is displayed is a fundamental feature for the production of post-GWAS summary statistics and plots using other programs. The output from the analysis is saved in a text file, the name of which is specified by the user. Each parameter analysed is recorded in a list under a header row that specifies the values in each column. Script 4.8, shows the contents of an example output file for five SNPs. The analysis output displayed is from using the Cox PH model. Table 4.5, contains a detailed description of all the output file content. The output is different depending on the statistical model used for analysis. The SNP identifier number, chromosome number, the base pair position of each SNP, and the *p*-value are the only four columns needed to produce a Manhattan plot.

```
InputName rsid Chr Pos EA NonEA CoefValue HR SE LowerCI UpperCI
    Waldpv LRTpv ModLRTpv EAF MAF Infoscore
SNP1 rs1234 10 13380 C G -10.01334 4.49E-05 42.63179 0 8.6952052E+31
     0.814341993 0.80691279370237 2.73440577082171E-29 0.99997639
    2.361E-05 -1
SNP2 rs1984 10 16154 C T -276.22756 0 184.139489 0 5.919171E+36
    0.133594002 0.114161788733577 8.16041572989629E-30 0.99997555
    2.445E-05 0.00111351
```

```
4 SNP3 rs4233 10 17544 T C -20.69194 0 32.044426 0 1.8063706E+18
    0.516986901 0.505106899627898 2.25935260263007E-29 0.99960455
    0.00039545 0.00291701
5 SNP4 rs5264 10 17599 T C -0.2082 0.814684 0.48563323 0.3140749
    2.11031612 0.672998575666 0.674152996 2.58006621541E-29
    0.63879319 0.36120681 0.01458189
6 SNP5 rs1229 10 18235 C A -31.0332 0 42.72901 0 7.789281203E+22
    0.46764205427 0.456629741423332 2.1393434820166E-29 0.99928541
    0.00071459 0.00131644
```

Script 4.8: SurvivalGWAS_SV text file output. Example output for five SNPs analysed using a Cox PH model.

Output header	**Description**
InputName	Variable name (can be the SNP ID, covariate or interaction name).
rsid	The unique SNP identifier for each SNP analysed.
Chr	User-specified chromosome number.
Pos	Base pair position of the SNP.
EA	Reference allele.
NonEA	Alternative allele.
CoefValue	Effect size estimate. Log-relative hazard (Cox model) or log-relative change in time (Weibull).
HR	Hazard ratio.
AF	Acceleration factor (Weibull only).
SE	Standard error of coefficient value.
LowerCI	Lower 95% confidence interval for HR (Cox model only).
UpperCI	Upper 95% confidence interval for HR (Cox model only).
Waldpv	Wald test *p*-value.
LRTpv	Likelihood ratio test *p*-value (Cox model only).
ModLRTpv	Model likelihood ratio test.
zscorestat	Score test statistic for *p*-value calculation (Weibull only).
p-value	*p*-value from z-statistic (Weibull only).

Table 4.5 continued from previous page

EAF	Effect/major allele frequency.
MAF	Minor allele frequency.
Infoscore	IMPUTE info measure of imputation quality. The info score metric takes the value 1 if all genotypes are completely certain and a value of 0 if the genotype probabilities for each sample are completely uncertain.
Shape	The shape parameter estimation of the survival distribution (Weibull only).

Table 4.5: SurvivalGWAS_SV output file variable headers and corresponding description.

4.5.8 Installation Guide

The software can be downloaded from the University of Liverpool, Statistical Genetics and Pharmacogenetics Research Group software page: https://www.liverpool.ac.uk/translational-medicine/research/statistical-genetics/survival-gwas-sv/. As well as the download, there is a full description of the software, a sample shell script and a frequently asked questions (FAQ) section. To download SurvivalGWAS_SV, navigate to the webpage URL and follow the instructions provided on the webpage. The downloaded folder contains the executable file and all the .NET framework .dll files needed to run and distribute the software. This software is also bound by the GNU General Public License, version 3 (GPL-3.0) with no restrictions to use, edit and distribute the software. The reasoning behind creating software like this is for the sole purpose of working collectively to advance scientific research in the areas of genetics, mathematics and bioinformatics.

4.6 Simulation Study

To demonstrate the utility of SurvivalGWAS_SV, we simulated genotype data using the software HAPGEN2 (Su et al. 2011), based on European ancestry individuals from the HapMap3 (Altshuler et al. 2010) reference panel. Approximately 1.5 million SNPs were simulated across 22 chromosomes for 1000 individuals. One SNP (rs12425539) on chromosome 12 was selected as the causal variant, to be used to generate TTE data. This SNP was selected because it represents common variation with a MAF of 0.317. We generated the TTE data using the power calculator software "SurvivalGWAS_-Power" (software described in detail in Chapter 3), which simulated the survival time and censoring indicator for each individual, for this single replicate of genotype data at the causal SNP.

A treatment covariate (binary) was also simulated for each individual using a binomial distribution. The active treatment and the placebo were divided evenly (1:1) between the 1000 individuals. All four datasets were simulated with right censoring occurring randomly within the sample using an exponential distribution:

1. Events under a PH assumption, with an additive effect of 0.6 in the log-hazard ratio for each copy of the minor allele. 579 censored observations.

2. Events under a PH assumption, with an additive effect of 0.6 in the log-hazard ratio for each copy of the minor allele, treatment effect of 0.4 and interaction effect of 0.5. 670 censored observations.

3. Events under an accelerated failure time (AFT) model, with an additive effect of 0.8 in the log-hazard ratio for each copy of the minor allele. 758 censored observations.

4. Events under an AFT model, with an additive effect of 0.8 in the log-hazard ratio for each copy of the minor allele, treatment effect of 0.4 and interaction effect of 0.5. 534 censored observations.

The PH data were simulated within SurvivalGWAS_Power, using the description in Section 3.4.3. G_i becomes the genotype of individual i at SNP rs12425539. $\mathbf{x_i}$ is the treatment covariate.

$$T_i = e^{2+\phi_s G_i + \phi_{\mathbf{x}} \mathbf{x_i} + \phi_\gamma G_i \mathbf{x_i} + Q\epsilon} \quad \text{(Eq. 4.1)}$$

The AFT data were simulated using the model described in Eq. 4.1, where the error distribution is represented as $\epsilon = Weibull(shape = 2, scale = 20)$ and $Q = 1/2$. For many distributions like the Weibull, an additional parameter Q is used to scale the error distribution. The parameters ϕ_s and ϕ_x are the effect on log-hazard of the effect allele at the SNP, and the treatment effect, respectively, and ϕ_γ is the interaction effect between the SNP and treatment.

The primary aim of the simulation study was to test whether our causal SNP on chromosome 12 can be identified as associated with our outcome using SurvivalGWAS_SV. Along with testing for associations with just a SNP effect, we investigated whether there is a SNP-treatment interaction effect present adjusting for treatment and SNP main effects. The statistical methodology choice was determined by the data, i.e. datasets (1) and (2) were analysed using the Cox PH model, whereas datasets (3) and (4) were analysed using the Weibull regression model. Only the SNP term was included in the analysis models for analysing datasets (1) and (3), while for datasets (2) and (4), SNP, treatment and interaction terms were included in the analysis models. After analysis, the number of SNPs was reduced by removing SNPs with a MAF < 0.01. This MAF threshold was used to eliminate rare variants for which there is minimal power to detect association and is considered a standard procedure in GWAS quality control. The secondary aim was to observe the efficiency of the software, measuring the length of time it took for completion of the analysis from submission of commands to output using several cores. Figures 4.3 and 4.4 present the results from the Cox PH model depicted by Manhattan[4], and quantile-quantile (QQ) plots for dataset (1). The Cox PH analysis was able to detect the causal SNP association,

[4]A Manhattan plot named after its resemblance to the Manhattan skyline, is a graphical representation of $-log_{10}$ p-values corresponding to chromosome position. It helps visualise and identify associated SNPs in the presence of millions of individual test results.

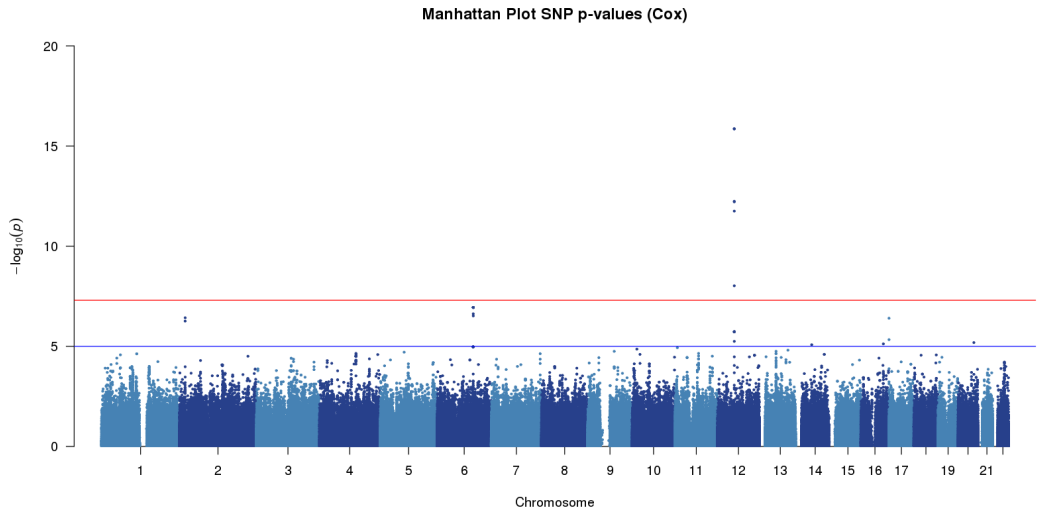


Figure 4.3: Manhattan plot of Cox PH analysis SNP p-values. Red line represents genome-wide significance threshold 5×10^{-8}. The blue line represents suggestive significance line. Each point represents a SNP. The two shades of blue used for SNPs is to distinguish between the chromosome boundaries.

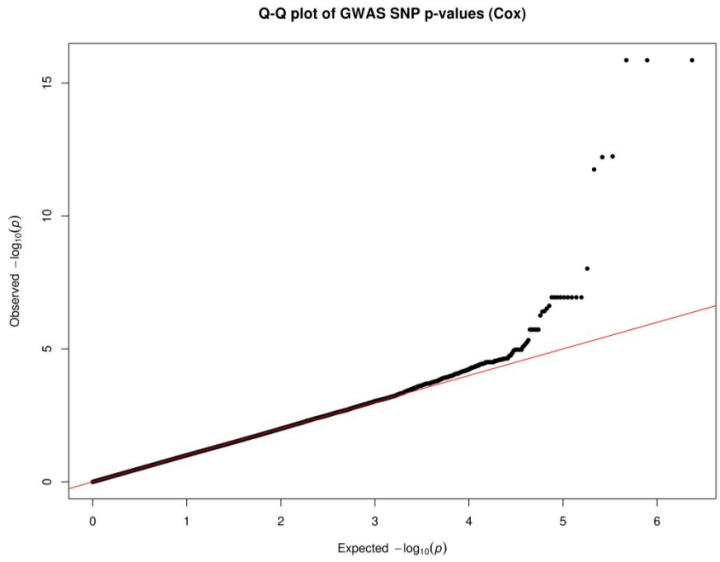


Figure 4.4: QQ-plot: Cox PH analysis of each SNP. Observed $-log_{10}$ p–values are plot against the expected $-log_{10}$ p–values.

identifying SNPs to be genome-wide significant in the data simulated using the PH model. The causal SNP (rs12425539) was the lead associated SNP from the analysis $\left(\beta_S = -0.579, HR = 0.560, SE(HR) = 0.070, P_{Wald} = 1.38 \times 10^{-16}\right)$. The true value of the simulated effect was $\phi_s = -0.6$.

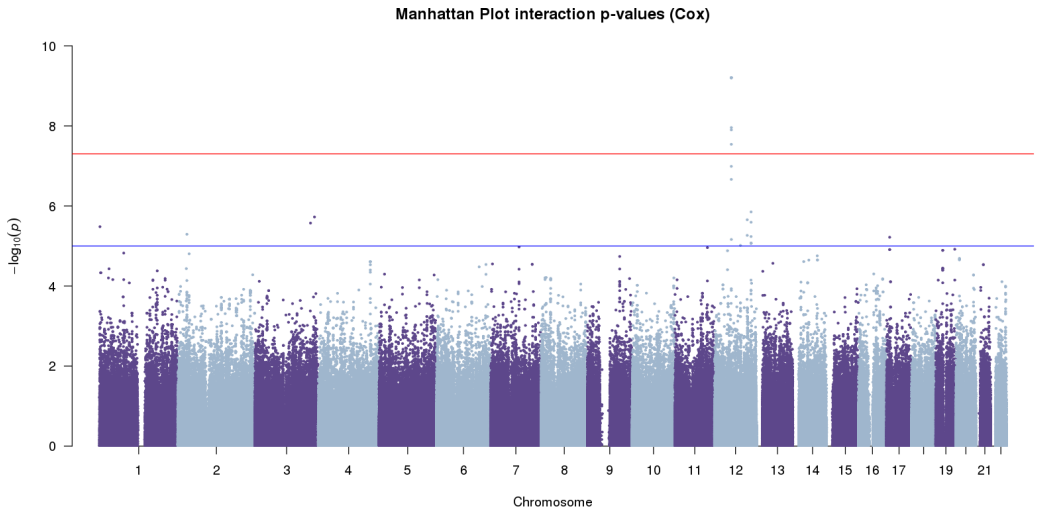


Figure 4.5: Manhattan plot of Cox PH analysis SNP-treatment interaction p-values. Red line represents genome-wide significance threshold 5×10^{-8}. The blue line represents suggestive significance line. Each point represents a SNP-treatment interaction. The two colours (purple and grey) are used to distinguish between the chromosome boundaries.

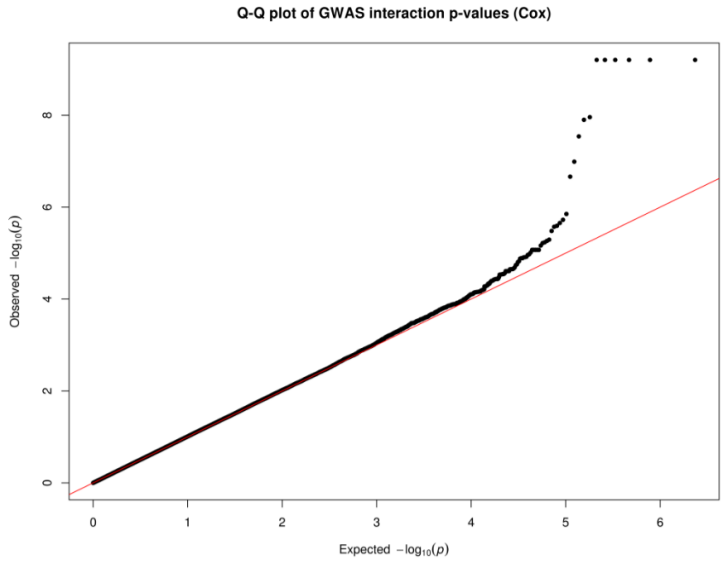


Figure 4.6: QQ-plot: Cox PH interaction analysis for each SNP-treatment interaction. Observed $-log_{10}$ p–values are plot against the expected $-log_{10}$ p–values.

Figures 4.5 and 4.6 depict the interaction analysis (2 degrees of freedom joint association LRT) for dataset (2), simulated using the PH model. The joint association effect comparing the alternative and null models was found to be genome-wide significant at SNP rs12425539 and a number of other SNPs at the same locus. Figures 4.7 and 4.8 represent the results from analysing the datasets simulated using the accelerated failure

time assumption. Figure 4.7 shows us that the Weibull regression analysis identified the association between the causal SNP and TTE outcome. This result was also observed for Figures 4.9 and 4.10, which illustrate the output from the joint association interaction analysis, indicating that the Weibull regression model was able to detect the interaction effect in dataset (4) at the causal SNP.

This simulation study has demonstrated the use of SurvivalGWAS_SV on a variety of different TTE datasets. SurvivalGWAS_SV has shown the ability to analyse large-scale genetic data, allowing for treatment covariate and SNP-treatment interaction effects. Under all settings, the causal SNP was identified to be statistically significant at genome-wide significance $\left(p < 5 \times 10^{-8}\right)$.

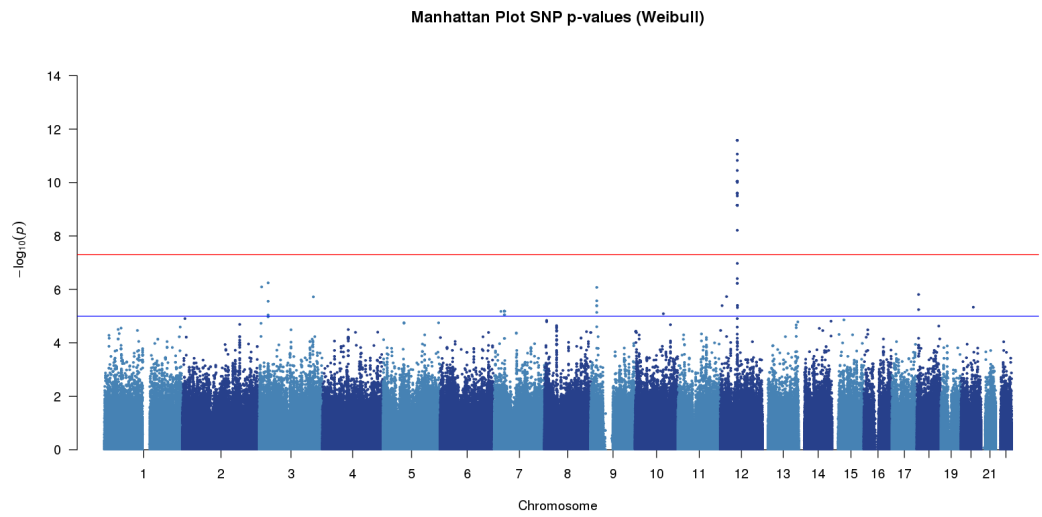


Figure 4.7: Manhattan plot of Weibull regression analysis SNP p-values. Red line represents genome-wide significance threshold 5×10^{-8}. The blue line represents suggestive significance line. Each point represents a SNP. The two shades of blue used for SNPs is to distinguish between the chromosome boundaries.

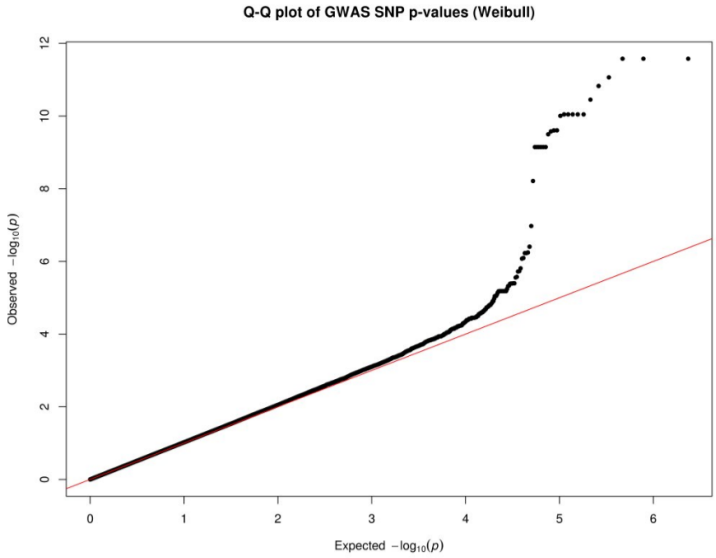


Figure 4.8: QQ-plot: Weibull-regression analysis of each SNP. Observed $-log_{10}$ p–values are plot against the expected $-log_{10}$ p–values.

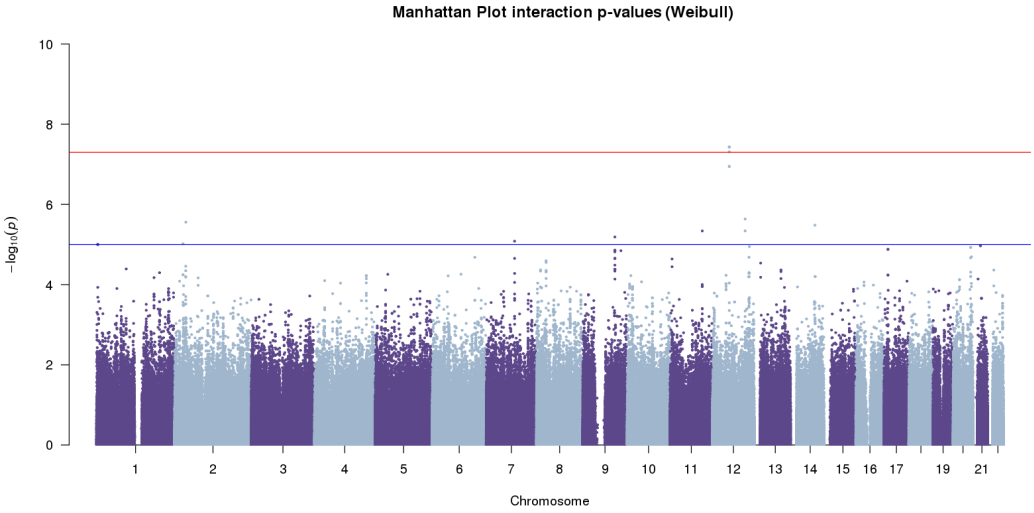


Figure 4.9: Manhattan plot of Weibull regression analysis SNP-treatment interaction p-values. Red line represents genome-wide significance threshold 5×10^{-8}. The blue line represents suggestive significance line. Each point represents a SNP-treatment interaction. The two colours (purple and grey) are used to distinguish between the chromosome boundaries.

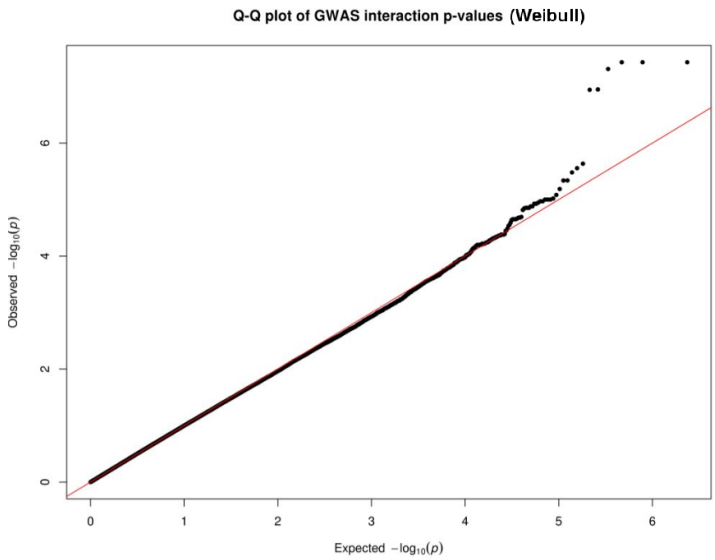


Figure 4.10: QQ-plot: Weibull-regression interaction analysis of each SNP-treatment interaction. Observed $-log_{10}$ p–values are plot against the expected $-log_{10}$ p–values.

4.6.1 Performance

The entire simulation study analysis was run using eight computer nodes (64 cores). Each node consisted of an HP Proliant DL170h G6 server, 2 Intel Xeon(R) E5520 2.27GHz quad-core CPUs, 36 GB memory and 1 TB of local storage. Running the single SNP analysis of ≈1.5 million SNPs across 22 chromosomes for 1000 individuals with no additional covariates took ≈4.5 hours to complete using the Cox PH model. Running the same analysis using the Weibull regression model took ≈3 hours to complete. The more covariates added to the analysis and the addition of an interaction, the longer the computational runtime. Each additional covariate took approximately an extra 0.00275 seconds for each SNP analysed.

The Weibull regression analysis runtime varies greatly; this is due to the convergence criteria of the Newton-Raphson method used for estimation of all parameters. Runtime is also dependent on missing values within the sample file and whether or not the genotype file is compressed. Ultimately, cluster specifications and the size of data files are the most influential factors affecting the speed of the software. As highlighted above, the only caveat is that Mono is used to compile the code on Linux O/S which

can reduce the performance. New compilers have become available recently such as .NET Core run using Visual Studio Code. This new framework could be beneficial for the future of the software.

4.7 Discussion

SurvivalGWAS_SV is the first analytics software capable of applying a range of survival analysis methods to genome-wide data, with appropriate handling of imputed genotypes. SurvivalGWAS_SV is compatible with HPC clusters, thereby allowing an application to large-scale GWAS datasets efficiently and effectively, without incurring memory issues. SurvivalGWAS_SV has the potential to enable discovery of genetic biomarkers of patient response to treatment for a range of complex human diseases and will offer opportunities for patient stratification according to predicted benefit or risk of treatment, allowing personalisation of therapeutic intervention.

Despite the great benefit in which GWAS single variant analyses have helped identify SNP-phenotype associations, much of the genetic heritability[5] cannot be explained. It has been suggested that rare variants, which are typically filtered out of GWAS analyses, might contribute to this missing heritability. This filtering of rare variants is undertaken because they would have insufficient power to detect associations using single variant tests. However, over the last seven years, methodology has been developed and implemented in software specifically to analyse rare variants collectively within sets, genes or other functional units. The current state of the methodology will be explored further in the context of TTE outcomes in the next chapter.

[5]The proportion of phenotypic variation within a trait occurring due to genetic variation. A value is usually derived by comparing the trait correlations/relatedness in individuals.

CHAPTER 5

RARE VARIANT ASSOCIATION STUDIES FOR TIME-TO-EVENT OUTCOMES

5.1 Overview

Methodology and software for the analysis of common variants within genome-wide association studies (GWAS) have been comprehensively developed and applied to a range of different endpoints, including survival data (see Chapter 4). In doing so, this has led to a vast number of loci identified for a variety of complex traits and diseases. As of 1 June 2017, the National Human Genome Research Institute's GWAS Catalog (MacArthur et al. 2017) reports 9346 single nucleotide polymorphism (SNP) - trait associations. All 9346 associations are significant at a p-value $\leq 5.0 \times 10^{-8}$ for a number of different phenotypes that include disease status, physiological measurements and response to drug. However, GWAS have been designed to identify common variant associations, which typically have modest effects, and together account for a small proportion of the genetic variance of the underlying trait (Manolio et al. 2009). It has been suggested that rare genetic variants may account for the "missing heritability" of human traits since they are not well covered through traditional GWAS, even after supplementation by high-density imputation.

An alternative approach to GWAS genotyping of common SNPs is studying potential genetic factors associated with complex traits through whole-genome and -exome sequencing. These sequencing solutions offer many advantages such as: i) direct investigation of low-frequency and rare variants that are not accessible through GWAS genotyping arrays, even after imputation; ii) through the use of annotation, rare variants can be aggregated into distinct groups based on similar criteria, e.g. position (gene or regulatory region), molecular effects, pathways or other biological sets. The groups are analysed using aggregate tests of association, which offer substantial gains in power

over single-variant tests.

One of the greatest challenges is designing software for the efficient analysis of rare variants because of the added complexity of the statistical models that need to be considered. As established earlier in Chapter 4 in the context of the analysis of common variants, there is also a lack of software analysing rare variants with time-to-event (TTE) outcomes.

5.1.1 Objectives

The focus of this chapter is to develop novel statistical methods and computationally efficient software for rare variant association study (RVAS) analysis of TTE outcomes, which can address the scale and complexity of sequence data. Specifically, the objectives are to:

1. Briefly review the role of rare genetic variants in common and complex human diseases.

2. Review the literature comparing and developing gene-based rare variant tests of association, while considering how to apply these to survival data.

3. Gather evidence on the application of rare variant tests for a range of outcomes within published studies.

4. Develop novel statistical methods that can be applied to RVAS with complex TTE outcomes.

5. Construct computationally efficient and freely available software for the implementation of the proposed methodology.

6. Demonstrate the utility of the proposed software through application to simulated data based on exome-array study data.

5.2 Evaluating Gene-based Analysis Methodology

Traditional methodologies for the analysis of GWAS lack power for detecting rare variant associations (Moutsianas et al. 2015). Instead, multiple rare variants are most often jointly analysed within "functional units", such as genes, exons or pathways. Aggregate or "gene-based" analyses are typically performed by assessing the association with rare variant "burden" or "dispersion" within the functional unit, focussed on the joint effects of multiple variants. These classes of tests each have their benefits and limitations, dependent on the underlying genetic architecture of the trait under investigation. Moutsianas & Morris (2014) provide a comprehensive overview of available rare variant tests and a discussion of architectures in which there is an advantage of using one method over another. In brief, burden tests focus on the effect of the mean number of minor alleles across variants in a gene. Burden tests are most effective when the direction of effects of all variants on the outcome in a given gene or functional unit is the same. On the other hand, dispersion tests focus on the effect of the variance in the number of minor alleles across variants in a gene and are most effective when variants have different direction of effects.

The literature, to date, has focused primarily on binary and quantitative traits (Moutsianas et al. 2015, Lee et al. 2014, Santorico & Hendricks 2016), and is expanding rapidly with extensions of existing approaches becoming available over the last few years (Wu et al. 2015, Ionita-Laza et al. 2013). The reason for this expansion of methodology is to account for the combination of different characteristics of variants that influence the power to detect associations. The power of aggregate tests for rare variants depends on: i) the number of variants per gene; ii) the proportion of variants (rare and common) in a gene that are causal; iii) the magnitude of effect of each variant; iv) the direction of effects for variants, i.e. risk or protective[1]; and v) the region length.

Appropriate approaches for variant aggregation while considering these five factors is imperative, because only a small fraction of variants within a gene may affect gene

[1] A variant can act to decrease or increase the trait or disease risk.

function, and therefore methods for classifying variants as to their effect on function need to be appropriately designed in the future. It is also essential that when variants have a different function, they are appropriately weighted for effect size because there could be one or many that are highly penetrant. If variants affect gene function in different directions, then inappropriately combining these into a set, may partially cancel out their effects. Definitions for combining variants within a gene or functional unit should be established before analysis. Methods for grouping *p*-values together after analysis have also been proposed (Lin et al. 2014). Currently, no method is uniformly most powerful to detect all rare-variant associations, therefore the consensus is to apply many if not all of the available tests to understand if there is an agreement between findings.

Chen et al. (2014) have previously introduced methodology behind the implementation of the general burden test (BT) and sequence kernel association test (SKAT) within a survival analysis framework. The SKAT is a dispersion test described in detail by Ionita-Laza et al. (2013) for binary traits. Chen et al. (2014) presented simulations evaluating the power of BTs and SKATs for TTE outcomes, considering only the Cox proportional hazards (PH) model, comparing the likelihood ratio test (LRT) with the score statistic. This paper was important in the development of new methodology for analysing rare variant associations and TTE outcomes, however coming to the same conclusions regarding statistical power being dependent on the underlying genetic architecture as reported for binary and quantitative traits. This paper provides a foundation for further model development and implementation within a computational analysis tool.

5.3 A Review of Rare Variant Association Studies

RVAS have been an actively researched area over the last few years, in the hope of uncovering the unaccounted genetic variance of human traits and diseases. Early studies (Wagner 2013) suffered from a lack of power to detect the effects of rare variants due to the genotyping chips available. Solutions through imputation from high-density

reference panels had been proposed and executed (Mägi et al. 2012) with success. However, the motivation that was once about costly re-sequencing studies has been overturned due to the substantial decrease in the cost of next-generation sequencing (NGS) experiments, allowing a more in-depth look at rare variants. This comprehensive inspection of rare variants is mostly performed using exome sequencing whereas whole genome sequencing is still considered expensive.

In the field of pharmacogenetics, in particular, there is evidence of a substantial contribution of rare coding variants to drug metabolism and transport (Gordon et al. 2014, Legge et al. 2017). These variants could have equally large effects on an individual's ability to metabolise certain drugs, resulting in response to treatment. de With et al. (2017) provide an overview of genetic studies conducted for clozapine[2], providing details of studies that have analysed rare-coding variants and their findings. However, no studies were identified from our literature search for pharmacogenetic TTE studies analysing rare coding variants. To date, very few RVAS have been undertaken regarding TTE outcomes, even in the context of pharmacogenetics. The few that exist inform us on how rare variants are analysed within TTE settings and the methodology and software currently employed to carry out the task.

Gaastra et al. (2016) sought to identify rare variants associated with amyotrophic lateral sclerosis (ALS) survival through a pre-screening of candidate survival genes found through previous evidence highlighted in the literature. A total of 50 samples were collected from the UK National DNA bank for motor neurone disease research. A second set of individuals from the Netherlands (n = 459) for replication of findings was also analysed. A single gene from three candidates was found to be associated, which contained a variant associated with longer survival and another with shorter survival. The statistical analysis was carried out by dichotomising patients into short and long survival groups and applying the sequence kernel association test (SKAT) (Wu et al. 2014) comparing the rare variation dispersion in the two groups. This analysis was conducted in variant association tool (VAT) and the R package '*SKAT*' (http://

[2]Clozapine is an anti-psychotic drug used to treat patients with treatment-resistant schizophrenia.

cran.r-project.org/web/packages/SKAT/). PLINK/SEQ (https://atgu.mgh.harvard.edu/plinkseq) was then used to identify the individual rare single nucleotide variations from the gene-based signals. The study produced a lack of evidence to suggest any associations in the replication set of samples from the Dutch cohort. They concluded with discussing the possible reasons for the lack of replication. First, the small sample size ($n = 50$), limits the statistical power, resulting in false positive associations. Second, rare variants are more likely to be population specific. The inclusion of only candidate survival genes (not looking at the entire genome) will have limited the discovery process. Furthermore, dichotomising the outcome may have reduced the power to detect true associations if present. As established in Chapter 2 of this thesis, dichotomising the survival outcome results in a loss of power to detect associations when compared to a TTE analysis.

Another ALS study (Pang et al. 2017), conducted within the Chinese population, had investigated the burden of rare variants in known ALS genes influencing survival for familial and sporadic ALS. A sample size of 8 patients were in the familial ALS group and a sample size of 46 in the sporadic ALS group. Kaplan-Meier curves were used to measure the effect of the rare variant burden on the outcomes time to ventilator and death. Alongside this, the Cox PH model was used to estimate the hazard ratio (HR) of the rare variant burden on survival. The rare variant association tests were only performed under a case-control design using RVTESTS (Zhan et al. 2016) and KGGSeq-integrated SKAT package (Ionita-Laza et al. 2013). The findings of the study were very promising as the burden of rare variants affected the survival probability of patients. The study found that patients with two or more rare variants had shorter survival compared to those with one or none. This study highlights the need for computational tools which can implement a combination of survival and rare variant analysis methodology.

Mackelprang et al. (2017) conducted a whole-genome sequencing study to identify variants associated with increased risk of acquiring HIV-1. The discovery stage used a

case-control approach to compare variation in genic regions. The method used was a logistic regression analysis with burden scoring developed by Morris & Zeggini (2010). The software implementing this method was not stated. In the replication stage of the analysis, the investigators tested the association between time to seroconversion and candidate regions from the initial discovery stage. *P*-values and HRs were estimated using the Cox model with an aggregate risk scoring based on whether or not an individual carried a minor allele for any of the rare variants.

As mentioned in the preceding chapters, TTE outcomes are commonly investigated in cancer studies, and rare-variant association is no different. Winham et al. (2016) undertook an exome-wide association study to investigate whether rare coding variants have an association with epithelial ovarian cancer survival. This study was undertaken because of the lack of associations reported for common genetic variants. The gene-level analysis was conducted using both the BT and SKAT within a Cox PH model, which was developed by Chen et al. (2014). The R package '*seqMeta*' (Voorman et al. 2017) was used to compute the gene-level meta-analysis, which implements the R '*survival*' package (Therneau 2015) functions.

These studies are informative about the role and benefits of rare variant association analyses with TTE outcomes. The examples discussed in this section all suffer from very low numbers of individuals in the sample, and adequate sample size is needed especially for detecting rare variant associations (Wagner 2013). Obtaining these numbers of subjects may be difficult for some phenotypes, such as treatment responses, however in pharmacogenetic studies we expect larger effect sizes than complex traits. The majority of the studies discussed above concluded with stating that more powerful studies should be undertaken, providing a sense that their findings can be advanced. The lack of software availability for detecting associations between rare variants and TTE outcomes led most investigators to using case-control designs with application through existing software and methodology. However, the impact of dichotomising the outcome on the statistical power to detect associations was not discussed in any of

the papers. This reiterates the emphasis throughout this thesis, that there is a need for bespoke methodology and software for TTE outcomes in genetic association studies.

5.4 Rare Variant Analysis Computational Tools

Software implementing rare variant tests for a variety of binary and quantitative traits, such as EPACTS (http://genome.sph.umich.edu/wiki/EPACTS), GRANVIL (Mägi et al. 2011) and VAT (Wang, Peng & Leal 2014), are well developed to handle genome- or exome-wide data from sequencing or array genotyping. However, these computational tools can only be applied to TTE data after dichotomising the outcome. As demonstrated in Chapter 2, this will result in a loss of information and power to detect associations. There are many implementations of rare variant association tests, with none considered to be optimal. Consequently, programs such as EPACTS and VAT are considered to be sophisticated programs because they can perform multiple tests, giving users' flexibility especially for omnibus[3] tests of association or adaptive tests such as the SKAT-O (Lee et al. 2012), which is a combination of the BT and SKAT.

Due to the demand for analysing more complex phenotypes, other computational tools exist such as '*RVFam*' developed by Chen & Yang (2016), a rare variant analysis software package that has a survival modelling element. This R package is used for rare variant association analysis of family data. The details of the package are that it is designed to analyse survival traits using Cox PH with shared frailty in each family by calling the 'coxph' function of the '*survival*' package in R. Although it does not offer a wider range of analyses, the software is useful for family study data, which is beyond the scope of this thesis.

To address the need for rare variant analysis of TTE outcomes, we have developed the rareSurvival analysis tool implemented using C# and run on Linux, Windows or Mac OSX operating systems (O/S). rareSurvival is capable of handling the scale and complexity of whole-genome sequence data offering support using analysis via the BT

[3]Omnibus tests refer to the application of multiple or a combination of many aggregate tests of association. For example, combining the p-values from both the SKAT and Burden tests (Derkach et al. 2013).

with optional weighting methods (discussed in Section 5.2) within a Cox PH or Weibull regression model framework.

5.5 rareSurvival

5.5.1 Implementation

Based on the same algorithmic framework as its companion program, SurvivalGWAS_-SV, rareSurvival has been developed for the analysis of rare-variants with TTE outcomes. rareSurvival has been developed using C# and is compatible with Windows and Linux or Mac O/S through Mono (http://www.mono-project.com/download/).

5.5.2 User Interface

rareSurvival is a console application that can be run from Linux, Windows and Mac OS X operating system terminals. Users can interactively apply the software on a command line interface or submit a script of commands to high-performance computing (HPC) capabilities, allowing more efficient analysis depending on the users' available resources. The user can specify batches of the data file to analyse using many computer cores, where each core can run the analysis process concurrently.

5.5.3 Inputs

Initially, the user is required to specify three data files that will be read into the program. The first file must be a genotype file which contains the SNP probabilities (imputed or typed) or a variant call format (VCF) file of sequence-based genotypes (v4.1 and v4.2 have been tested). The second file should be a sample file (.sample or .txt) which contains all the covariate and phenotype information for each individual. The final file to specify is the gene list file which can be in one of two different formats. The gene list file is a text file which can contain: (i) the gene name, chromosome, base pair start position and base pair end position, which should be given the file extension .pos (see

Script 5.1); or (ii) the gene name, chromosome and the rsid of each variant in a row list separated by a single space, which should be given the extension name .list. The software can read in genotype files that are compressed, either with a .zip or .gz file extension. GZIP should be used to compress genotype files into the .gz format. All files should have content separated by a single space or tab; otherwise, errors will occur.

```
GeneA 1 11873 11909
GeneB 1 12851 13469
GeneC 1 13671 14425
GeneD 1 14362 16765
```

Script 5.1: Gene list file (.pos) contents for four genes.

In addition, the user needs to specify details about variables to include in their analysis model such as covariates, while also specifying the censoring indicator and survival time. Furthermore, they need to specify the range of functional units to be analysed, based on the number of lines in the gene list file. Since the program will search through the genotype file for every line in the gene list file, the analysis process can be slow; however, this can be avoided if the user reads in smaller batches of the input files separated by chromosomes.

The user must enter the details of the survival analysis method to use and the rare variant test separately (see Table 5.1). If the data have not already been filtered for MAF or imputation quality, then a MAF and info-score threshold can be specified. The default values for filtering out variants based on MAF and info-score are greater than 0.05 and 0.9, respectively. Finally, the name of the output file should be specified for which the analysis output will be saved. If the user adjusts for covariates within the analysis model but does not require the covariate results in the output file, then there is a print option that when specified only outputs the gene analysis results.

5.5.4 Algorithms and Validation

rareSurvival ensures that the data are processed through a comprehensive validation procedure. The gene list file is first assessed by the software determining how many analyses need to be distributed between threads. Once this has been determined the first line is read in by rareSurvival. After this the genotype file is read in one line at a time, rareSurvival will convert the genotype probabilities into a SNP coded under an additive dosage model for the minor allele (see Eq. 1.2). This step is only carried out for probabilities; if hard genotype calls or dosages are present within a VCF file, then these are directly read in. The hard genotype calls from VCF files can be in the form $0/0, 0/1, 1/1$ or $0 \mid 0, 0 \mid 1, 1 \mid 1$.

Two alleles are listed for each variant in the genotype file. The first is assumed to be the reference allele, the second to be the alternate or non-reference allele. However, due to imputation, it is not unusual for the ordering to be different. The software identifies which is the major allele and which is the minor allele by calculating the MAF. In cases where the hard calls, probabilities or dosages reflect that the major and minor allele are swapped due to allele frequency calculation then the genotype values are flipped. For example, a variant has reference allele coded A and alternative coded as T. The genotypes in the file is represented as [AA, AT, TT], with genotype calls [0, 1, 2]. The alternative allele frequency is calculated to be 0.99. Therefore the alternative and reference allele are switched to [2, 1, 0] because the reference allele is the minor allele in this case. This is a trivial process, especially when handling dosage information. Users are advised to check that data is in the correct format before being read in using rareSurvival.

Each SNP, in turn, will be matched to the genes base pair position in the gene list file or the rsid of the variant, if present in the list. If the variant does not match the specifications in the gene list file, then the program moves to the next variant. However, if the variant is to be included, it is then stored in a new data frame of included SNPs. The program will determine how many SNPs are included once it has scanned the

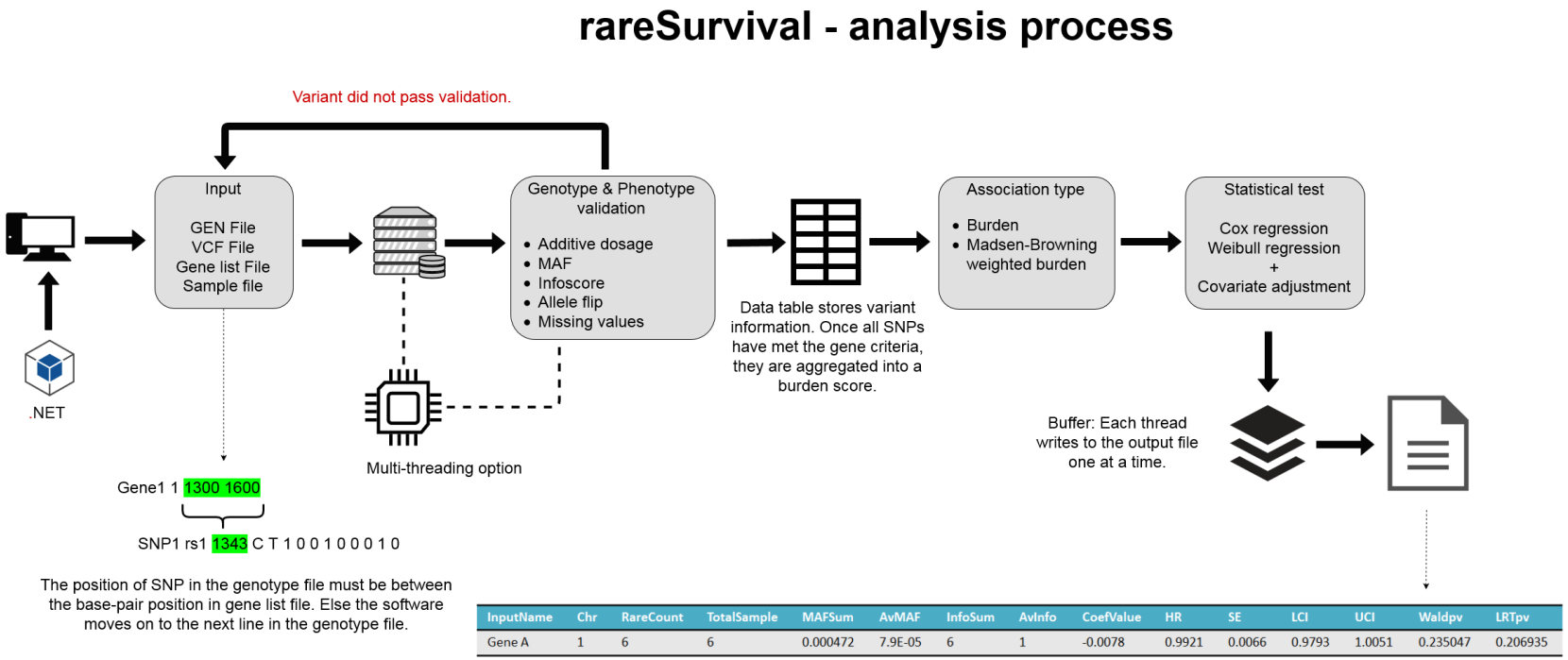
entire genotype file or until it has added at least one variant and then identified that the next variant to be read in is outside of the specified genomic region. All SNPs within the region will be taken forward to the analysis stage. The group of SNPs within the gene are analysed, and the software starts the process again reading in the next line of the gene list file. rareSurvival throws exemptions whenever the user has specified an incorrect command or states a heading name that cannot be found in the data files, at which point the program will exit the application, and it will need re-submission of the task. The program also handles missing values (coded as "NA") within the sample file. If an individual has missing values for survival time or a covariate used in the model, then the individual is removed from the analysis with their corresponding variant information. Figure 5.1 provides a detailed work-flow of rareSurvival from reading the data to analysis output.

5.5.5 Statistical Methodology

As described in Section 5.2, there is no single gene-based analysis model which has the greatest statistical power to detect associations amongst all genetic architectures. This observation prompted an initial implementation of the BT of association and Madsen-Browning (Madsen & Browning 2009) weighted BT within TTE regression models. Chen et al. (2014) provided the framework to implement the burden test into rareSurvival. In contrast to BTs, the SKAT, which was also described by Chen et al. (2014) was not implemented because it is very computationally demanding and mathematically cumbersome therefore the BT was more of a straightforward implementation. rareSurvival offers analysis using the Cox PH and Weibull regression models. For each, a BT with or without Madsen-Browning weights is implemented.

Burden Test within a Cox and Weibull Regression Model

The BT is an aggregated test of association for the genetic burden score. The genetic burden score is the accumulation of minor alleles at rare variants. There are many different types of BT, the simplest of which is the sum of all genotype dosages within



InputName	Chr	RareCount	TotalSample	MAFSum	AvMAF	InfoSum	AvInfo	CoefValue	HR	SE	LCI	UCI	Waldpv	LRTpv
Gene A	1	6	6	0.000472	7.9E-05	6	1	-0.0078	0.9921	0.0066	0.9793	1.0051	0.235047	0.206935

Figure 5.1: rareSurvival quality control and analysis pipeline.

a genomic region with the addition of applying an indicator of whether at least one rare variant exists within a genomic region. These tests are most powerful when all of the variants in a region are causal with effects in the same direction. This burden is introduced into the Cox PH model or Weibull regression model as a predictor in the same way as a coded SNP genotype in a single variant analysis. Coefficients are then derived with the corresponding p-value using a Wald test (see Section 1.3) or the LRT. Under this model, let B_i denote the genetic burden of a given gene or functional unit for individual i, given by:

$$B_i = \sum_{j=1}^{m} G_{ij} w_j \qquad \text{(Eq. 5.1)}$$

In this expression the indicator variable,

$$w_j = \begin{cases} 1, & \text{if j is included} \\ 0, & \text{otherwise} \end{cases}$$

G_{ij} represents the genotype coding of the i'th individual at the j'th rare variant in the gene. m is the total number of variants within/mapping to a given functional unit. Each genotype at a variant of interest is coded under an additive dosage model for the j'th rare variant.

Using Eq. 1.3, we again denote the TTE for the i'th individual by t_i, and a vector of covariates $\hat{\mathbf{x}}_\mathbf{i}$. Under the assumption of PH, we can express the hazard of the event occurring at some time t for the i'th individual by:

$$h_i(t) = h_0(t) e^{\beta_G B_i + \hat{\beta}_\mathbf{x} \hat{\mathbf{x}}_\mathbf{i}} \qquad \text{(Eq. 5.2)}$$

In this model, the parameters β_G and $\hat{\beta}_\mathbf{x}$ correspond to the effect on log-hazard of the burden score, and the vector of covariates, respectively. β_G can now be interpreted as the log-HR per minor allele across variants in the functional unit. The baseline hazard is represented by $h_0(t)$.

Using Eq. 1.7, the likelihood function can be adapted for a more general Weibull regression model accounting for right censored data analysing the cumulative effects of

a functional unit. This is carried out by replacing β_s (the effect of a single SNP) with β_G and the SNP G_i with the burden score B_i.

Madsen-Browning Weighting within a Cox and Weibull Regression Model

The issue with gene-based tests of association is the imperfect classification of the effects of variants when most likely a small percentage of variants within a functional unit are causal. Under the unit-weighting model of the classical BT, all rare variants included in the aggregation are assumed to have the same magnitude of effect on the phenotype, as well as the same direction. However, this cannot be the case for all variants within a functional unit, and simple aggregation methods may incorrectly combine these into a set. This method can result in partly cancelling out variant effects that can lead to false-negative associations. This makes the task of finding those variants that are driving the association much more difficult. Typically, variants are excluded from the analysis on the basis of MAF and annotation. Alternatively, w_j can be used to incorporate weights according to allele frequency.

Different functional variants will have a different magnitude of effect and should be weighted appropriately in a gene-based test. There are many different types of weighting schemes for a BT, such as, the Madsen-Browning approach that gives greater weight to variants of lower frequency. They proposed a weight that increases the impact on the phenotype for lower frequency variants, ensuring that larger effect sizes are given to variants with very small MAF. This approach is based on the expectation that rarer variants are more likely to have arisen from recent mutation events, and therefore selection will have had less opportunity to have removed variants with large detrimental effects from the population. The weight w_j in Eq. 5.1 now becomes: $w_j = \frac{1}{\sqrt{(q_j(1-q_j))}}$, where q_j is the MAF of the j'th variant.

5.5.6 Usage Commands

```
1 $ mono raresurvival.exe -gf= -sf= -mf= -threads= -t= -c= -cov= -i= -
```

```
chr= -maf= -info= -lstart= -lstop= -method= -rm= -p= -o=
```

Script 5.2: rareSurvival command line example without defined parameters.

Command	**Description**
-gf=	This specifies the genotype file. Typically a .gen, .vcf, .gen.gz.
-sf=	This specifies the sample file (.sample).
-mf=	This specifies the the gene list file. (.pos or .list).
-threads=	Specifies the number of threads. On a multi-core system, multiple, threads can execute tasks in parallel, with each core executing a different thread or multiple threads.
-t=	This specifies the time to event (column heading name) in the sample file.
-c=	This specifies the censoring indicator/outcome in the sample file.
-cov=	This specifies the covariates to adjust for in the model. Each one separated by a comma (,). e.g. -cov=cov1,cov2,cov3. Note: Categorical variables need to be converted to binary as software only assumes continuous or binary covariates.
-lstart=	This specifies the line in the gene list file at which the start, position of analysis will occur. Used to break large files into small batches for parallel computing.
-lstop=	This specifies the line in the gene list file at which the end position, of analysis. will occur. Typically the number of lines is equal to the number of genomic regions in the file.
-chr=	This specifies the chromosome number to be output in the text file.
-p=	Enter "onlygene" if only the results from the gene analysis are to be output.
-m=	This specifies choice of regression model. This is either "cox" for the Cox PH model or "weibull" for the parametric Weibull regression model.

Table 5.1 continued from previous page

-rm=	Specifies the choice of rare variant analysis method. This is "burden" for the BT with unit weighting or "mbweight" for the Madsen-Browning weighted BT.
-maf=	Specifies the MAF threshold for inclusion of variants in the analysis.
-info=	Specifies the info-score threshold for inclusion of variants in the analysis.
-o=	This specifies the name of the file for output to be saved in. e.g. name.txt
-help	Outputs a full list of commands and usage help.

Table 5.1: List of commands available in rareSurvival and their corresponding usage description.

Assuming all data files and software are in the same folder, the command line shown in Script 5.3 is an example for interactive submission in a Linux terminal. The command is specifying an analysis of 100 genes on chromosome 6 with one additional covariate using a BT in a Cox PH model. Filtering based on SNPs with MAF less than equal to 0.01 is also specified.

```
$ mono raresurvival.exe -threads=4 -gf=data.vcf -sf=data.sample -mf=
    list.txt -t=event_times -c=outcome -cov=covariate1 -chr=6 -lstart
    =0 -lstop=100 -maf=0.01 -m=cox -rm=burden -p=onlygene -o=output.
    txt
```

Script 5.3: rareSurvival command line example with dummy input parameters.

As demonstrated in Script 5.2 and 5.3, each command is separated by a space and begins with '-' and ends with '=' before specifying the option, identical to the commands for SurvivalGWAS_SV (Section 4.5). The user can specify the exact location of the data files and where the output file will be saved. e.g. /DIRECTORY/DATA/output.txt. Script 5.4 is an example of a shell script (.sh) to distribute the analyses using rareSur-

vival, between 10 computer cores within a Linux cluster, using a Sun-Grid engine batch system.

```
#!/bin/bash
#$ -o stdout
#$ -e stderr

DIRECTORY=/rareSurvival #Location of software and data
str1=0 #Start position in gene list file
str=100 #Number of genes/lines in gene list file
no_of_jobs=10 #Number of cores
inc=`expr \( $str - $str1 \) \/ $no_of_jobs` #Increment

#SGE_TASK_ID takes values 1:no_of_jobs
nstart=`expr \( $SGE_TASK_ID - 1 \) \* $inc`
nstop=`expr $nstart + $inc - 1`
mono $DIRECTORY/raresurvival.exe -threads=4 -gf=$DIRECTORY/data.vcf
    -sf=$DIRECTORY/data.sample -mf=$DIRECTORY/list.txt -t=event_times
     -c=outcome -cov=covariate1 -chr=6 -lstart=0 -lstop=100 -maf=0.01
     -info=0.8 -m=cox -rm=burden -p=onlygene -o=$DIRECTORY/output${
    SGE_TASK_ID}.txt
```

Script 5.4: Shell script that runs rareSurvival on a HPC cluster. Comments are highlighted in green.

To submit the script file, there are many different commands dependent on the cluster manager used. Script 5.5 demonstrates submission using the "qsub" command. The script and command can be easily changed for alternative cluster workload managers, such as SLURM, replace "SGE" with "SLURM_ARRAY" in Script 5.4 and submit using the command line shown in Script 5.6. The concatenation of files should be carried out using the same command as Script 4.7 in Section 4.5.

```
$ qsub -t 1:10 example.sh
```

Script 5.5: Multitple core submission example using ‘qsub’.

```
$ sbatch --array=1-10 example.sh
```

Script 5.6: Multitple core submission example using 'sbatch'.

5.5.7 Output

The output from the analysis is saved in a text file specified by the user. Details of the file contents are summarised in Table 5.2. In addition to the standard HR and *p*-value output, key output such as the total number of rare variants within a genomic region provide a value for which the user can filter out single markers. This process is done because it is of interest to only include the tests of groups with more than one variant. Often the total MAF of all rare variants in the gene or the mean MAF per variant in the gene is of interest, as this can also be used as an additional tool for filtering. Script 5.7 is a sample of an output text file showing analysis of four genes.

```
InputName/Gene Chr RareCount TotalSample MAFSum AvMAF InfoSum AvInfo
    CoefValue HR SE LowerCI UpperCI Waldpv LRTpv ModLRTpv
GeneA 10 1 1 0.000236 0.000236 1 1 -0.31479947 0.72993523
    621.36388324 0 Infinity 0.999595770477929 NA 0.548746029322796
GeneB 10 1 1 0.007075 0.007075 1 1 0.02452142 1.02482455 0.01844898
    0.98843072 1.06255839 0.183799127698855 NA 0.204347599530058
GeneC 10 8 8 0.000943 0.000118 1 0.125 0.01261971 1.01269967
    0.00599925 1.00086215 1.0246772 0.0354179036648269 NA
    0.0500864643631872
GeneD 10 2 2 0.000236 0.000118 1 0.5 -0.01065098 0.98940554
    0.01685544 0.95725453 1.0226364 0.527451151265893 NA
    0.506230074265232
```

Script 5.7: rareSurvival text file output. Example output for four genes analysed using a BT within a Cox PH model.

Output header	Description
InputName/Gene	Variable name (can be the gene name or covariate).
Chr	User-specified chromosome number.
RareCount	Total number of rare variants in gene.
TotalSample	Total number of variants (common and rare) in gene.
MAFSum	Sum of rare variant MAFs in gene.
AvMAF	Mean of rare variant MAFs in gene.
InfoSum	Sum of rare variant info-scores in gene.
AvMAF	Mean of rare variant info-scores in gene.
CoefValue	Coefficient estimated value.
HR	Hazard ratio of accumulation of minor alleles.
AF	Acceleration factor (Weibull only).
SE	Standard error of coefficient value.
LowerCI	Lower 95% confidence interval for HR (Cox model only).
UpperCI	Upper 95% confidence interval for HR (Cox model only).
Waldpv	Wald test *p*-value.
LRTpv	Likelihood ratio test *p*-value (Cox model only).
ModLRTpv	Model likelihood ratio test.
Shape	The shape parameter estimation of the survival distribution (Weibull only).

Table 5.2: rareSurvival output file variable headers and corresponding description.

5.5.8 System and Installation Guide

The software can be downloaded from the University of Liverpool, Statistical Genetics and Pharmacogenomics Research Group software page: https://www.liverpool.ac.uk/translational-medicine/research/statistical-genetics/raresurvival/. As well as the download, there is a full description of the software, and a sample shell script. rareSurvival is publicly available and can be obtained by visiting the web-link, and then by following the instructions on the web page. The downloaded folder contains the executable file and all the .NET framework .dll files needed to run and distribute the software. This software is also bound by the GNU General Public License, version 3 (GPL-3.0) with

no restrictions to use, edit and distribute the software.

5.6 Simulation Study

To demonstrate the features of the software and to provide an evaluation of its efficiency, we conducted simulations based on Illumina Exome Array genotype data (primarily coding variants) obtained from 2,120 individuals from two cohorts of elderly Swedish individuals.

5.6.1 Background

The two cohorts were from the Prospective Investigation of Vasculature in Uppsala Seniors (PIVUS) (http://www.medsci.uu.se/pivus/) and the Uppsala Longitudinal Study of Adult Men (ULSAM) (http://www.pubcare.uu.se/ulsam/Database). The cohorts were typed for the Illumina HumanExome-12 v1 Beadchip at the Wellcome Trust Centre for Human Genetics, University of Oxford. Genotype calling and quality control were undertaken at the University of Oxford. In short, genotypes were initially called with the Illumina GenomeStudio GENCALL software (Guo et al. 2014). Poor quality samples were excluded based on call rate (< 98%), extreme heterozygosity, outlying numbers of singletons[4], sex discordance and ancestry outliers. Poor quality variants were excluded based on call rate (< 95%), and extreme deviation from Hardy-Weinberg equilibrium $\left(p < 10^{-4}\right)$. Missing genotypes were then recalled using zCALL (Goldstein et al. 2012). For this final called genotype set, poor quality samples and variants were excluded on the basis of call rate (< 99%).

5.6.2 Simulation models

The phenotype information was simulated using SurvivalGWAS_Power. Two datasets were simulated, each with a randomly selected gene containing causal variants.

[4]Singletons are the rarest of rare variants across the genome. A singleton is unique because it is found in a single individual in a population.

Dataset 1

The first gene selected as causal was *FNDC1* found on chromosome 6, which included a total of twenty-five rare variants. Fifteen variants were selected at random to be causal, simulating survival times for each individual using the number of minor alleles carried by an individual (Eq. 5.1). All variants had an effect size of 0.6 and with the same direction of effect. In this expression, v_{ij} represents the individual i, genotype at the j'th causal variant.

$$T_i = Weibull\left(1, 10e^{-\left(\sum_{j=1}^{15} 0.6 v_{ij}\right)}\right) \qquad \text{(Eq. 5.1)}$$

Dataset 2

The second gene selected as causal was *OR5B17* found on chromosome 11, which has a total of 4 rare variants. Survival times were simulated based on all 4 causal variants but with varying effect sizes $\kappa_j = [0.1, 0.2, 0.6, 0.15]$ with the same direction of effects.

$$T_i = Weibull\left(1, 10e^{-\left(\sum_{j=1}^{4} \kappa_\mathbf{j} v_{ij}\right)}\right) \qquad \text{(Eq. 5.2)}$$

Both model settings are suited towards analysis using a BT, as expressed in the comparative literature (Santorico & Hendricks 2016, Lee et al. 2014). However, there are subtle differences in both datasets, which suggest that they are not entirely ideal for the BT. The first setting has just over half the variants in the gene as causal, and in the second setting the magnitude of effects of the causal variants is different.

Censoring was randomly simulated using an exponential distribution with scale parameter 10. If the censoring time is less than the individual's event time, then the individual is censored. There were 1008 censored observations in the *FNDC1* dataset and 867 censored observations in the *OR5B17*. Two additional covariates were simulated for the sole purpose of collecting information regarding analysis completion times. These were treatment, a binary covariate simulated using a Bernoulli distribution, and age of individual, a continuous covariate simulated with a Uniform distribution generating

values between 20 and 50. More information on the variants used to simulate data can be found in Table 5.3.

The objective of the simulation study was not to make another comparison of methods, but rather testing the features of the software and computational efficiency. Interest lies with rare coding variants, so all variants with MAF >5% in the dataset were filtered out. Variants were assigned to genes and analysed based on base pair position using the gene list file which links coding variants (defined by a strict annotation) to genes. Analysis was completed on a total of 73952 SNPs over 12432 genomic regions across 23 chromosomes for 2120 individuals. Hereafter, filtering for genes that contained at least two rare variants was performed before visualising the output.

Gene	Unique variant identity number	Ref/alt allele	Position	MAF	Function	Simulated effect size
FNDC1	-	C/T	159,636,159	0.000235	Unknown	0.6
	rs186515442	G/A	159,646,577	0.000235	Missense	0.6
	rs200758408	G/A	159,659,662	0.001650	Missense	0.6
	rs61746218	C/T	159,667,972	0.002594	Missense	0.6
	rs200171920	A/T	159,672,511	0.000235	Missense	0.6
	rs180849332	C/T	159,687,181	0.001650	Missense	0.6
	-	G/A	159,692,377	0.002830	Unknown	0.6
	rs200925962	A/G	159,692,428	0.000943	Missense	0.6
	rs202080149	A/G	159,644,604	0.000235	Missense	0.6
	rs202114028	G/A	159,653,261	0.000235	Missense	0.6
	rs199900169	G/A	159,636,039	0.000943	Missense	0.6
	rs201387402	A/G	159,644,575	0.000471	Missense	0.6
	-	A/G	159,653,612	0.0004717	Unknown	0.6
	rs7763726	A/G	159,670,100	0.020047	Missense	0.6
	rs186422799	G/A	159,653,921	0.006839	Missense	0.6

Table 5.3 continued from previous page

OR5B17	rs144440324	G/C	58,125,740	0.000235	Missense	0.1
	rs55810057	A/G	58,125,774	0.018632	Missense	0.2
	rs199650837	A/T	58,126,153	0.000943	Stop-gained	0.6
	rs140465731	C/T	58,126,340	0.000471	Missense	0.15

Table 5.3: List of causal variants used in simulation study. Stop-gained: leading to a gain of a stop codon. Missense: leading to an amino acid change. Abbreviations: MAF, minor allele frequency; Ref, reference; Alt, alternative.

5.6.3 Results

Dataset 1

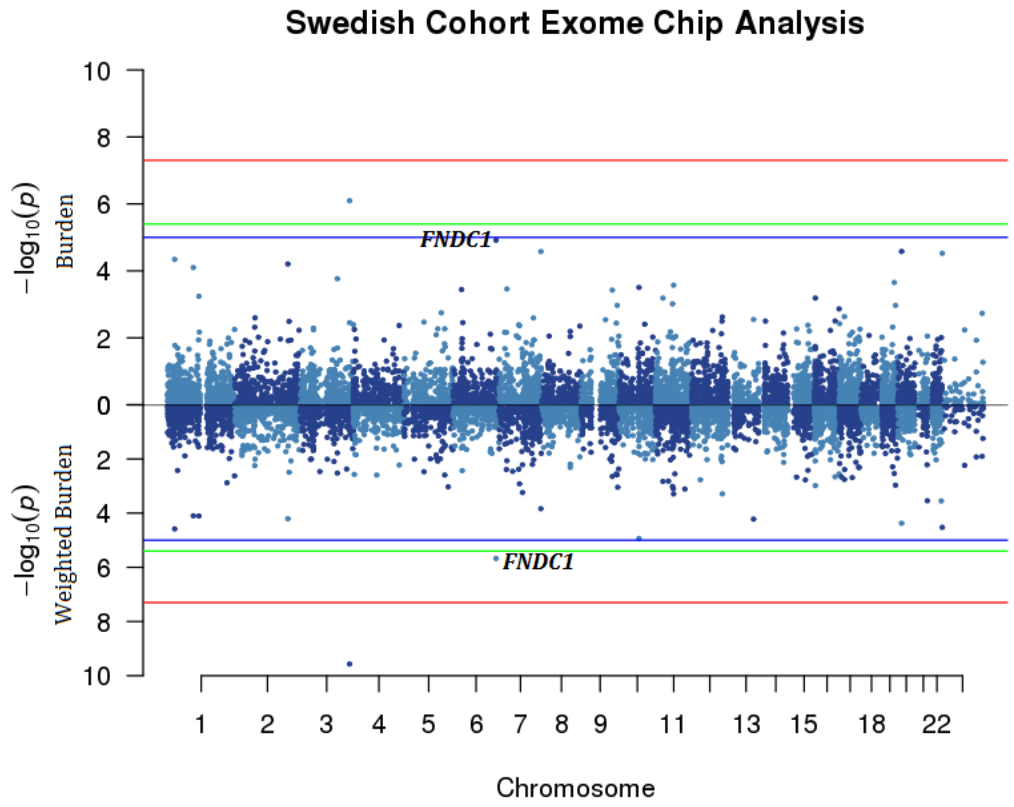


Figure 5.2: Mirrored Manhattan plot of dataset 1, comparing alternative rare variant tests within a Cox PH model. Red line: Genome-wide significance $\left(5 \times 10^{-8}\right)$, Blue line: Suggested significance $\left(1 \times 10^{-5}\right)$, Green line: Bonferroni corrected exome-wide significance $\left(4.0 \times 10^{-6}\right)$.

Figure 5.2 presents the results from the analysis of the first simulation study dataset using both a BT and Madsen-Browning weighted BT in a Cox PH model. The mirrored Manhattan plot shows us that both methods were successful in identifying the *FNDC1* gene with the causal variants. The *FNDC1* gene is highly associated with the phenotype using the weighted burden test at a Bonferroni corrected significance level for the total number of genes (12,432 genes) at 4.0×10^{-6} (represented in Figure 5.2 with a green line). However, the association signal found by the unweighted burden test was not significant at the corrected threshold but was close to the suggested significance line. The p-values presented represent that of the Wald test. Table 5.4 shows a summary of the top genes found by both rare variant tests, comparing estimation using the Wald

and LRT. There is close agreement between the p-values of all 4 tests for *FNDC1*, with the Wald test p-value from the Weighted BT as the most significantly associated.

Many other genes represented in Table 5.4 are significantly associated with the simulated TTE outcome. All of the genes have very low total MAF, which is indicative of a false positive association. The variants within these genes may be similar to the variants from the *FNDC1* gene in terms of the minor allele for each individual at each variant. Only the gene *HTR3D* is significantly associated at the Bonferroni corrected threshold of 4.0×10^{-6}. All associated genes other than *FNDC1*, have a smaller number of aggregated rare variants and it is likely that if most of the variants in these genes are similar to the causal variants within *FNDC1* then it is reasonable to assume that the genes would show strong association with the TTE outcome. Furthermore, the sample size for this dataset is relatively small. Therefore, the minor allele at a rare variant may, by chance, appear in an individual with an extreme trait value. The gene-based analyses, only need a few of these rare variants to generate a highly significant result, where the chance of such an outcome increases as the sample size decreases.

The accumulation of the simulated effect for dataset 1 was 0.6 across 15 variants. There were 25 variants in total within *FNDC1*, with 10 variants with unknown effects on outcome (i.e. not included in the simulation model). The average known estimate of the HR for the simulated burden would be $e^{(0.6\times15)/25} = 1.433329$. The estimated HR from the BT is 1.31861988. The HR is underestimated because of the inclusion of non-causal variants in the analysis.

Gene	CHR	Rare Count	MAF Total	$WaldP_{Burden}$	$LRTP_{Burden}$	$WaldP_{MB}$	$LRTP_{MB}$
ESYT3	3	10	0.000236	0.000171	8.53×10^{-5}	0.003013	9.93×10^{-4}
HTR3D	3	5	0.000236	8.01×10^{-7}	3.42×10^{-6}	2.70×10^{-10}	1.23×10^{-6}
FNDC1	**6**	**25**	**0.000943**	$\mathbf{1.22 \times 10^{-5}}$	$\mathbf{2.54 \times 10^{-5}}$	$\mathbf{2.13 \times 10^{-6}}$	$\mathbf{1.64 \times 10^{-5}}$
MTFR1L	1	3	0.000236	4.54×10^{-5}	0.001331	2.62×10^{-5}	0.001317
SLC44A3	1	2	0.000236	7.96×10^{-5}	0.006332	7.96×10^{-5}	0.006332
CCDC150	2	3	0.000472	6.23×10^{-5}	0.002874	6.23×10^{-5}	0.002874
INSIG1	7	3	0.000472	2.65×10^{-5}	0.001281	1.46×10^{-4}	0.003591
NDUFAF5	20	3	0.000708	2.63×10^{-5}	0.000718	4.23×10^{-5}	0.001449
ARSH	23	3	0.000236	2.99×10^{-5}	0.006049	2.99×10^{-5}	0.006052
TTC16	9	3	0.000708	0.003641	0.000935	0.002806	8.09×10^{-5}

Table 5.4: Table of significantly associated genes from *FNDC1* simulated data. Abbreviations: CHR, chromosome; LRT, likelihood ratio test; MB, Madsen-Browning.

Dataset 2

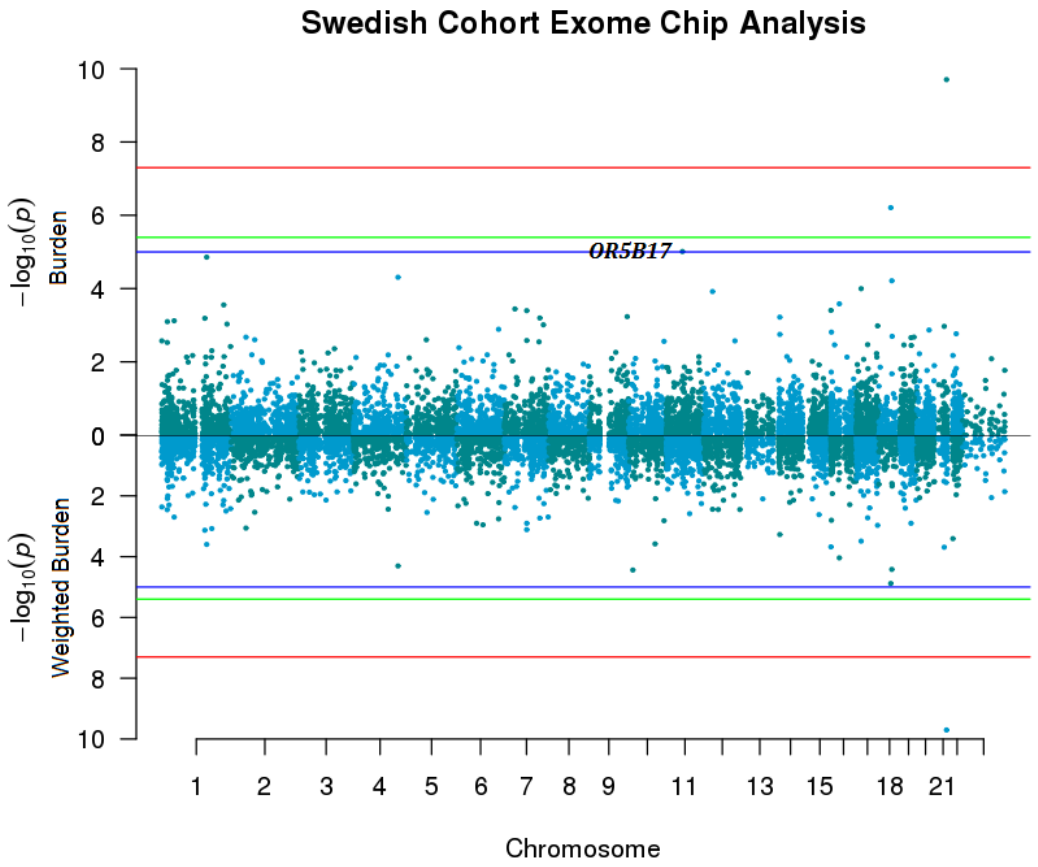


Figure 5.3: Mirrored Manhattan plot of dataset 2, comparing alternative rare variant tests within a Cox PH model. Red line: Genome-wide significance $\left(5 \times 10^{-8}\right)$, Blue line: Suggested significance $\left(1 \times 10^{-5}\right)$, Green line: Bonferroni corrected exome-wide significance $\left(4.0 \times 10^{-6}\right)$.

In contrast to the data simulated using the *FNDC1* gene, *OR5B17* contained a smaller set of SNPs with each variant contributing to the underlying association signal. The results of the *OR5B17* gene analysis was found to be inconsistent between the four tests. *OR5B17*, shows strong association when analysed using the classical BT, but not at exome-wide significance. The discrepancy between the *p*-values for each test could be due to the variety of low to intermediate effects for the causal variants the TTE outcome was simulated with.

Similarly to the results from the *FNDC1* dataset, Table 5.5 shows a number of significantly associated genes for the *OR5B17* simulated dataset. As mentioned earlier these

associations are most likely the result of the small sample size, very low MAF and similarities between the variants in those genes with the variants in the *OR5B17* gene for which the TTE outcome was simulated.

The accumulation of the simulated effect for dataset 2 was 0.2625 across 4 variants. There were 4 variants in total for the gene *OR5B17* all contributing to the effect on outcome within the simulation model. The average known estimate of the HR for the simulated burden would be $e^{0.2625} = 1.300176$. The estimated HR from the BT is 1.78928621. The average effect per minor allele for the BT, is inconsistent with the average simulated effect.

Gene	CHR	Rare Count	MAF Total	$WaldP_{Burden}$	$LRTP_{Burden}$	$WaldP_{MB}$	$LRTP_{MB}$
RSPH1	21	2	0.000236	1.98×10^{-10}	0.000142	1.98×10^{-10}	0.000142
HHIPL2	1	8	0.004481	0.000279	3.58×10^{-5}	0.007074	0.001859
OR5B17	**11**	**4**	**0.000472**	$\mathbf{9.75 \times 10^{-6}}$	$\mathbf{4.79 \times 10^{-5}}$	**0.025496**	**0.044330**
SLC14A1	18	4	0.000708	6.19×10^{-7}	0.000171	1.31×10^{-5}	0.000652
ITGA8	10	12	0.000236	0.117635	0.126302	3.63×10^{-5}	0.000383
WDR87	19	6	0.015566	0.002192	0.001404	0.001253	0.000225

Table 5.5: Table of significantly associated genes from *OR5B17* simulated data. Abbrevitions: CHR, chromosome; LRT, likelihood ratio test; MB, Madsen-Browning.

5.6.4 Performance

The entire analysis was run using 5 computer nodes (40 cores). Each node consisted of a HP Proliant DL170h G6 server, 2 Intel Xeon(R) E5520 2.27GHz quad-core CPUs, 36 GB memory and 1 TB of local storage. Running the complete gene-based analysis of each dataset using a Cox PH model with a BT and weighted BT under our simulation study setting with no additional covariates took ≈20 hours to complete (4 sets of analysis). The more covariates added to the analysis the longer the computational runtime. Computational runtime is highly dependent on sample size, missing values within the sample file, size (i.e. the number of variants) and compression status of the genotype file. The largest contributing factor is ultimately the specifications of the computing resource available to the user. Only the run times for the *FNDC1* simulated data are presented in Table 5.6 because both datasets have identical sample size and number of genes. The Madsen-Browning weighted BT effects the runtime minimally. Approximately a tenth of a second slower to complete the analysis of the *FNDC1* dataset compared to the classical BT.

Number of covariates	**Runtime**
0	324.198 minutes
Treatment (Binary)	378 minutes
Treatment & Age (Continuous)	445.2 minutes

Table 5.6: Computational runtime of the simulation study analysis using rareSurvival with and without additional covariates. Number of covariates including adjustment for each gene. Computational runtime of *FNDC1* simulated data analysis using BT within a Cox PH model.

5.7 Discussion

rareSurvival is a new analysis program for RVAS with TTE outcomes. The utility of rareSurvival has been demonstrated through analysis of exome-array simulated data using the novel methodology that combines aggregated tests of association within a

regression framework for survival outcomes. A caveat, as with most rare variant association software the analysis process is slow, though it can be increased with the separation of data into smaller sets, i.e. by chromosome, and linking each separated genotype file to its corresponding gene list file. rareSurvival is the first analytics software capable of applying a variety of survival analysis models together with gene-based tests of association to population-based RVAS data.

rareSurvival will play a key role in the discovery of novel genes associated with patient response to treatment for a range of complex human traits and diseases. Using rareSurvival for the combination of sequence-based genetic data with clinical data will help us better understand the genetic architecture of complex diseases, facilitating the translation of statistical results into practical solutions to advance disease prediction and treatment.

RVAS and GWAS should be considered as complementary to one another in the search for causal variants for a number of complex traits and diseases. Methodology and software development for RVAS together with the ongoing effort for GWAS will result in improved modelling of TTE outcomes. With the advancements towards NGS data and the pursuit of the analysis of more complex outcomes, future versions of rareSurvival will incorporate multifaceted survival models, together with the latest rare variant analysis methods, to account for non-PH, competing risks, variable variant effect sizes and directions. The use of both rareSurvival and SurvivalGWAS_SV are demonstrated through the analysis of the Pharmacogenetics of Acute Coronary Syndrome study data in the next chapter.

CHAPTER 6

PHARMACOGENETICS OF ACUTE CORONARY SYNDROME

The objective of this chapter was to apply SurvivalGWAS_SV and rareSurvival to the Pharmacogenetics of Acute Coronary Syndrome (PhACS) study data. The primary aim was to investigate the genetic basis of recurrence of coronary events and all-cause mortality following a primary acute coronary event. The secondary aim was to test the performance of the software through the application to this genome-wide association study (GWAS) of common and rare variants. Details of the study design, quality control (QC) procedure, exploratory analysis of covariates, association analyses and further investigation of significant variants and genes are provided.

6.1 Background

6.1.1 Cardiovascular Disease

Acute Coronary Syndrome (ACS) is classified as a cardiovascular disease (CVD) that occurs in situations where the blood supplied to the heart is abruptly blocked. Both heart attacks and myocardial infarction (MI) come under the term ACS, which is the leading cause of mortality worldwide (http://www.who.int/mediacentre/factsheets/fs310/en/), affecting more than 7 million people in the UK alone, of which an estimated 160,000 deaths occur each year (BHF 2017).

Many therapeutic options exist for the treatment of ACS, ranging from prescription drugs (such as aspirin, clopidogrel, statins and beta-blockers), through to surgical intervention, such as percutaneous coronary intervention (PCI) and coronary artery bypass grafting (CABG). There is significant inter-patient variability in response to cardiovascular drugs, and factors contributing to this include patient demographics

(e.g. age, gender) and medical history (e.g. smoking status) (Winham et al. 2015, Turner et al. 2015, El Desoky et al. 2006). More recently, researchers have searched for a genetic basis to this variability. Many studies are investigating this issue, which has generated some statistically significant findings: e.g. SNPs in *CYP2C19* for Clopidogrel (Dean 2015) and *ADRBI* for Beta-blockers (Shin & Johnson 2010). Franchini (2016) provides an overview of candidate gene studies and GWAS conducted for ACS.

6.1.2 Study Objective

The objective of PhACS was to investigate the association between genetic variants and treatment response, in terms of mortality, non-fatal myocardial infarction, non-fatal stroke, bleeding events and cardiac failure after an acute coronary event.

6.1.3 Study Design

PhACS is a UK prospective pharmacogenetic cohort study involving 1,470 patients who have had an ACS as their main diagnosis at hospital admission. Patients were then followed up prospectively for up to 48 months after discharge with patient demographics and relevant clinical factors recorded. Turner et al. (2017) describes the study design and cohort selection in detail.

6.1.4 Phenotypes

The time-to-event (TTE) outcome was defined as corresponding to the occurrence of any of the following events after hospital discharge (D/C): MI, stroke or cardiovascular death. The primary outcome was the time to any one of the following outcomes, whichever occurred first after baseline D/C to the end of the study: cardiovascular mortality, non-fatal MI or non-fatal stroke. If none of these events occurred during follow-up, patients were censored either at death for non-cardiovascular reasons or end of follow-up, whichever occurred first. The secondary TTE outcome was time to all-cause mortality after D/C. If mortality did not occur, patients were censored at the

end of follow-up. Throughout this chapter, time to cardiovascular event will be referred to as the primary outcome and time to all-cause mortality as the secondary outcome.

6.1.5 Genotype Data

Patients were genotyped using the Illumina Omni Express array. Furthermore, the design of the array adopted a comprehensive genotyping strategy to maximise the genome coverage, which involved using data from the HapMap (Altshuler et al. 2010), 1000 Genomes Project (Auton et al. 2015) and previous literature, to genotype all common functional variants and tagging SNPs, as well as rare variants ensuring that the diversity of the whole genome is accounted for.

6.1.6 Quality Control Procedure

Before the association analyses, a comprehensive QC procedure was initiated using the protocol outlined in Section 1.2.3. This procedure was carried out using the software PLINK (Purcell et al. 2007) and QCTOOL (http://www.well.ox.ac.uk/~gav/qctool_v2/documentation/alphabetical_options.html). The raw data based on all available samples were in PLINK binary (BED, BIM, FAM) format before being converted into GEN format. Genotype imputation was conducted using IMPUTE2 (Howie et al. 2012), and the 1000 Genomes Project (Auton et al. 2015) data.

6.2 Analysis Plan

The initial step of analysis was to identify possible significantly associated clinical risk factors for drug response. This selection method was undertaken using a stepwise Cox proportional hazard (PH) regression framework, implementing both forwards and backwards selection to identify a model of significant covariates, eliminating factors from the model which were no longer significant at a Wald p-value significance threshold of 0.05. This analysis was carried out for both the primary and secondary TTE outcomes.

Information on several non-genetic factors collected throughout the study are summarised into four categories; patient demographics, medical history, drugs given to patients from baseline hospital admission and other treatment (see Table 6.1). These non-genetic factors are often treated as confounders, though the role of a confounder should be further investigated via criteria of confounding (Greenland et al. 1999). The controlling for confounders is done to remove bias of the estimated association between the genetic factor and the outcome (Greenland et al. 1999, Shmueli 2010). A further note on the selection of covariates is in Shmueli (2010) who emphasises that the selection depends on the role of analysis: if it is predictive (or prognostic) vs explanatory (causal).

Factor Category	**Covariate**	**Description**
Patient Demographic	Site	Two sites: Liverpool & Blackpool.
	Age	Range from 26 to 94 years old.
	Sex	Male = 0, Female = 1.
	BMI	Body mass index.
Medical History	Hypertension	High blood pressure.
	Hyperlipidaemia	High levels of lipids e.g. cholesterol.
	CRF	Chronic renal failure.
	Diabetes mellitus	Type 1 and 2.
	Prior MI	Prior Myocardial Infarction (up to 5 years before admission).
	Past TIA or stroke	Prior Transient Ischaemic Attack with neurological deficit lasting less than 24 hours (up to 5 years before admission).
	Smoking status	Non-Smoker (0) - Never smoked. Current smoker (1) - within past 3 months. Previous-smoker (2) - stopped smoking for more than 3 months.

Table 6.1 continued from previous page

	Troponin[1] level	Troponin index raised or normal.
Drugs (D/C)	Aspirin	On treatment=1, if not = 0.
	Clopidogrel	On treatment=1, if not = 0.
	Beta blocker	On treatment=1, if not = 0.
	Statin	On treatment=1, if not = 0.
	ACEI	On treatment=1, if not = 0.
	Aldosterone antagonist	On treatment=1, if not = 0.
Other Treatment	PCI	Percutaneous coronary intervention.
	CABG	Coronary artery bypass grafting.
	Beta-blocker (A/D)	Beta-blocker taken before admission into hospital.
	ACEI (A/D)	Angiotensin-converting enzyme inhibitor.

Table 6.1: PhACS study clinical factor information. Abbreviations: PCI, percutaneous coronary intervention; MI, myocardial infarction; TIA, transient ischaemic attack; ACEI, angiotensin-converting enzyme inhibitor; D/C, after hospital discharge; A/D, before hospital admission.

Each significant ($p < 0.05$) factor from the stepwise Cox PH model was graphically visualised through Kaplan-Meier curves and tested for violation of the PH assumption through Schoenfeld residuals (Schoenfeld 1982) and diagnostic $-log(log(S(t))$ vs. $log(t)$ plots.

The association analysis involves testing for the association of each genetic variant passing QC with both TTE outcomes separately using a Cox PH model, assuming an

[1]Troponin is a protein that is released into the bloodstream during a heart attack (https://www.bhf.org.uk/heart-matters-magazine/medical/ask-the-experts/troponin).

additive dosage model. Adjustments are made for the significant clinical factors found using the stepwise regression model and principal components (PCs) to account for population structure (see Section 1.2.4) in the association analyses. The analyses are performed using SurvivalGWAS_SV.

Evaluating the evidence for association of each outcome with rare variants within each gene was conducted using a gene-based burden test (BT) within a Cox PH framework implemented in rareSurvival. Again, these analyses have been adjusted for non-genetic factors and PCs identified as significant. The gene-based analyses focussed on assessing rare variants using the gene classifications/annotations attained from the UCSC genome browser gene classification list (Kent et al. 2002). Each row in the list contains a transcript, which is a set of exons and a gene can contain multiple transcripts. The significance is based on a Bonferroni corrected threshold for the number of transcripts analysed. The threshold is $0.05/70663 = 7.07 \times 10^{-7}$.

For both single-variant and gene-based approaches, all analyses include all patients with complete phenotype and significant covariate information. The output of interest from the analyses is the p-value, which will be used to identify statistically significant genetic loci through the visualisation of Manhattan plots. All QC procedures and analyses were undertaken with a server comprising 8 computer nodes (64 cores). Each node consisted of an HP Proliant DL170h G6 server, 2 Intel Xeon(R) E5520 2.27GHz quad-core CPUs, 36 GB memory and 1 TB of local storage.

All significant SNPs found using the single-variant approach were further analysed through Kaplan-Meier curves in R (R Core Team 2013), to distinguish the survival between patients split by genotype group. Biological information used throughout Section 6.3, regarding gene function is provided by GeneCards, the human gene database (Safran et al. 2010) (http://www.genecards.org/) and UniProt, the central hub database for functional information on proteins (UniProt 2017). This information was used to link any of the significantly associated genes found from both the single-variant and rare-variant analyses to pathways and expression that may affect cardiovascular events

or mortality.

6.3 Results

6.3.1 SNP QC

500,387 SNPs with high missing genotype rate (> 5%) were removed from a total of 2,252,914 SNPs. A test for HWE was undertaken at each SNP using Fisher's exact test. A further 1,154,473 SNPs were removed due to deviation from HWE at the threshold $p < 0.0001$. Before checking for duplicated/related individuals, markers were removed using linkage disequilibrium (LD) based pruning, removing markers with high LD regions (LD > 0.2)

6.3.2 Sample QC

72 individuals were removed with low sample call rate (< 95%). Two individuals were removed due to an event time of 0 days recorded. These two patients were censored during the time between admission into hospital and discharge from the hospital.

Four individuals were removed due to duplication and two due to the relatedness between samples. This was calculated using the identity-by-state (IBS) matrix and identity by descent (IBD) coefficient. Gender checks against collected clinical data were undertaken, removing seven individuals that failed to match.

Population structure was investigated through principal components analysis (PCA) for the merged study data and Hapmap3 data populations: (i) Utah residents with Northern and Western European ancestry (CEU); (ii) Han Chinese in Beijing (CHB); (iii) Japanese in Tokyo (JPT); and (iv) Yoruba in Ibadan (YRI). 17 ethnic outliers were identified and excluded. PCA was conducted again without the HapMap3 data. The R package '*SNPRelate*' (Zheng et al. 2012) was used to perform the PCA. Figure 6.1 shows the amount of population variation explained by each of the first six PCs. It can be noted from Figure 6.1 that the first two PCs explain the majority of the population-

specific variation. The other PCs explain very little of the variance. Therefore only the first two PCs have been adjusted for in the association analyses. After applying

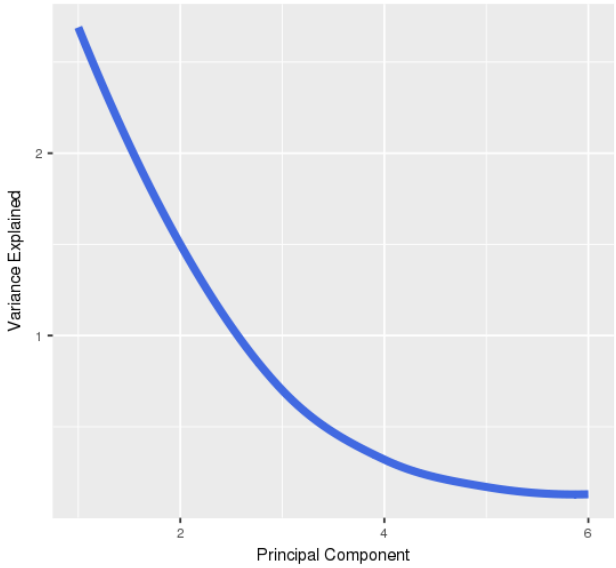


Figure 6.1: Proportion of variation explained by principal components. Calculated using the '*SNPRelate*' package.

sample and SNP quality control, 1367 samples and 598, 054 SNPs remained. The data were then imputed up to the 1000 Genomes Phase I reference panel for all ancestries (March 2012 release). Imputation was performed using software packages, SHAPEIT2 (O'Connell et al. 2014) and IMPUTE2.

After imputation QC the data were separated into variants with MAF $>$ 0.01 (low to common frequency variants) and MAF $<$ 0.01 (rare variants). A MAF of $<$ 0.01 was chosen because imputation using the 1000 Genomes reference panel produces a high proportion of low-frequency variants. Both sets were filtered for variants with an imputation accuracy (info-score) $>$ 0.4. A total of 8,788,380 SNPs remained for the single-variant analysis and 5,253,064 SNPs over 70663 functional units (defined as gene boundaries using the UCSC genome browser) remained for the rare-variant analysis.

6.3.3 Exploratory Analysis of Clinical Covariates

Demographics and clinical risk factors were tested for association with the primary and secondary outcomes using a Cox PH framework. The proportion of right-censored observations for the primary outcome was 86.24% (188 patient events and 1179 patient non-events). Median event free survival was 1658 days.

	HR	95% LCI	95% UCI	P_{Wald}
Age	1.042	1.027	1.056	9.02×10^{-9}
Prior MI	1.921	1.417	2.603	2.57×10^{-5}
ACEI A/D	1.579	1.172	2.128	0.00268
Aldosterone antagonist	2.047	1.255	3.339	0.00408
CRF	1.544	1.034	2.307	0.03387

Table 6.2: PhACS: Primary outcome stepwise regression model output. Abbreviations: HR, hazard ratio; LCI, lower confidence interval for hazard ratio; UCI, upper confidence interval for hazard ratio; P_{Wald}, the p-value calculated using the Wald test; CRF, chronic renal failure; MI, myocardial infaction; ACEI, angiotensin-converting enzyme inhibitor; A/D, before admission.

Several risk factors were identified to be significantly associated with the primary outcome as displayed in Table 6.2. Older patients are more likely to have an occurrence of a cardiovascular event (HR = 1.042, 95% CI = 1.027-1.056). Patients with medical history such as previous heart attack and chronic renal failure have an increased hazard of a recurrent cardiovascular event than those without this medical history (Prior MI:HR = 1.921, 95% CI = 1.417-2.603; CRF:HR = 1.544, 95% CI = 1.034-2.307). Use of ACE inhibitors prior to admission was associated with an increased hazard of an event compared with patients not taking the drug (HR = 1.579, 95% CI = 1.172-2.128). Even though this is a treatment for high blood pressure, they are often prescribed because the individual has previously had a heart attack. Therefore it is not unusual to see that individuals have an increased hazard of a cardiovascular event occurring when taking ACEI because they have had a prior MI which is also highly associated with the primary outcome. Patients using Aldosterone antagonist after hospital discharge are twice as likely to have an event (HR = 2.047, 95% CI = 1.255-3.339). A patient using a particular drug pre-admission or after hospital discharge may indicate the health condition of an

individual and will ultimately affect the patient's survival outcome. The secondary

	HR	95% LCI	95% UCI	P_{Wald}
Age	1.079	1.059	1.099	3.33×10^{-16}
Prior MI	1.564	1.095	2.236	0.01404
PCI	0.533	0.316	0.899	0.01845
CRF	1.698	1.099	2.623	0.01706
Hyperlipidaemia	1.654	1.134	2.423	0.00905
Statin	0.457	0.264	0.791	0.00512
Aspirin	0.565	0.346	0.923	0.02258

Table 6.3: PhACS: Secondary outcome stepwise regression model output. Abbreviations: HR, hazard ratio; LCI, lower confidence interval for hazard ratio; UCI, upper confidence interval for hazard ratio; P_{Wald}, the p-value calculated using the Wald test; CRF, chronic renal failure; MI, myocardial infarction; PCI, percutaneous coronary intervention.

outcome has a proportion of 90.4% censored observation (131 patient events compared to 1236 patient non-events). Many of the same clinical factors associated with the primary outcome are significantly associated with all-cause mortality (see Table 6.3), such as CRF, older age and previous history of heart attack. Patients that have had CRF or prior MI are at an increased hazard of mortality (Prior MI: HR = 1.564, 95% CI = 1.095-2.236; CRF: HR = 1.698, 95% CI = 1.099-2.623). Older patients are more likely to have an event that causes mortality (HR = 1.079, 95% CI = 1.059-1.099). PCI is associated with a reduced hazard of mortality, where half as many patients experience the event compared to those that have not had surgical intervention (HR = 0.533, 95% CI = 0.316-0.899). Patients with hyperlipidemia have 1.6 times the likelihood of mortality (HR = 1.654, 95% CI = 1.134-2.423). Patients who take statins or aspirin after discharge have a reduced hazard of mortality than those not taking these drugs (Statins: HR = 0.457, 95% CI = 0.264-0.791; Aspirin: HR = 0.565, 95% CI = 0.346-0.923).

6.3.4 Diagnostic Plots of Clinical Covariates

The Kaplan-Meier curve for prior MI shown in Figure 6.2 suggests event-free (cardiovascular event) survival amongst patients with no previous MI is significantly better

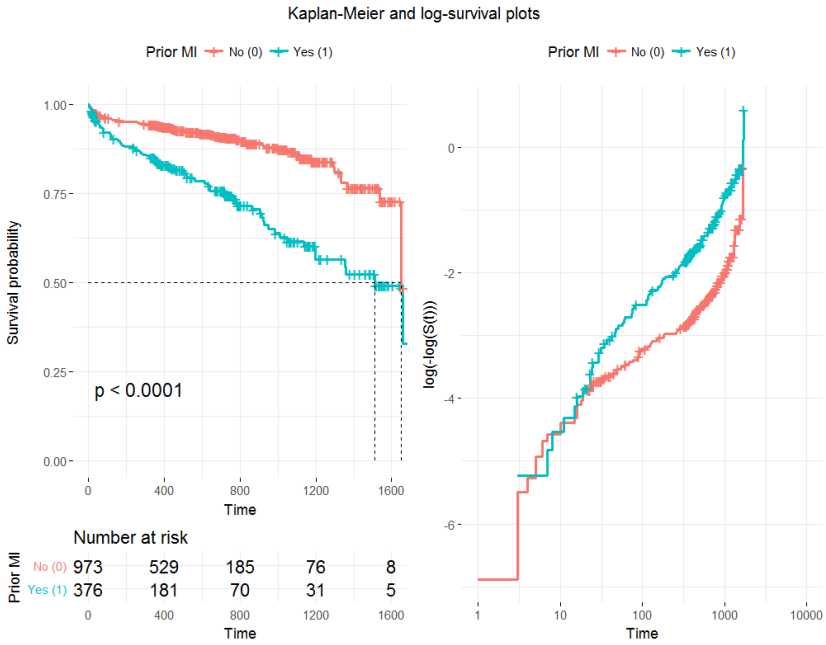


Figure 6.2: Prior Myocardial Infarction: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals for the primary outcome.

than those that had a prior MI. The diagnostic plot indicates that the PH assumption holds, however showing a slight deviation when the sample size within the risk set is low. This observation cannot be considered a violation of the PH assumption.

Use of ACE inhibitors pre-hospital admission is shown to increase the likelihood of a cardiovascular event (Figure A.1). The PH assumption holds for the majority of event times. Individuals taking aldosterone are depicted in Figure A.2 for having a significant reduction in survival. Median event-free survival is estimated to be at approximately 950 days for those on aldosterone and about 1680 days for those not. However, the sample size is small. The diagnostic plot is in keeping with the PH assumption. Figure A.3 compares those with and without CRF. The survival of the two groups is statistically significant, indicating that those with CRF are more likely to experience a cardiovascular event. The PH assumption holds for CRF.

Age is a continuous covariate, such that the proportionality assumption is assessed using Schoenfeld residuals as shown in Figure 6.3. From the top left plot of Figure 6.3, the residuals for age have distinctively no pattern with time. Consequently, it is not in violation of the PH assumption. Assessing all the significant binary covariates again

with the primary outcome, but through Schoenfeld residuals, all show a very minimal pattern with time, none of which form two distinct lines of residuals, therefore, PH hold for all non-genetic factors.

Kaplan-Meier plots for the secondary outcome are located in Appendix A, Figure A.4 shows that those with no history of prior MI have better survival than those that have had an MI. Figure A.5 tells us that having the PCI treatment intervention reduces the risk of mortality. Patients with hyperlipidemia have a minimal reduction in survival up to 800 days as shown in Figure A.6, after this the difference between the two groups is statistically significant. Figure A.7 suggests that patients with CRF have reduced survival compared to those without CRF. As Figure A.8 illustrates, the survival probability of patients taking aspirin after discharge significantly improved as opposed to those, not on aspirin. The majority of patients have been given aspirin after discharge. However, all are taking multiple treatments. Patients taking statins show an improvement in survival (see Figure A.9) over those not on statins, in the same way as aspirin; however, the difference is less statistically significant. All significant binary covariates for the secondary outcome, show parallel event times with no crossing hazards which supports the PH assumption (see Figure A.4 to A.9). Figure A.10 depicts Schoenfeld residual plots for age and all significant binary clinical covariates. All covariates show a random pattern with time indicating that the PH assumption holds.

6.3.5 Single Variant Association Analysis

All patients were analysed using a Cox PH model to estimate the association of each variant with the primary and secondary TTE outcomes. A Cox PH model was fitted to the number of copies of the minor allele at each SNP, significant covariates and PCs accounting for right censored data. A Wald test (see Section 1.3) was conducted to obtain the p-values.

Figure 6.4 summarises the p-values from the Cox PH for all SNPs. The GWAS of the

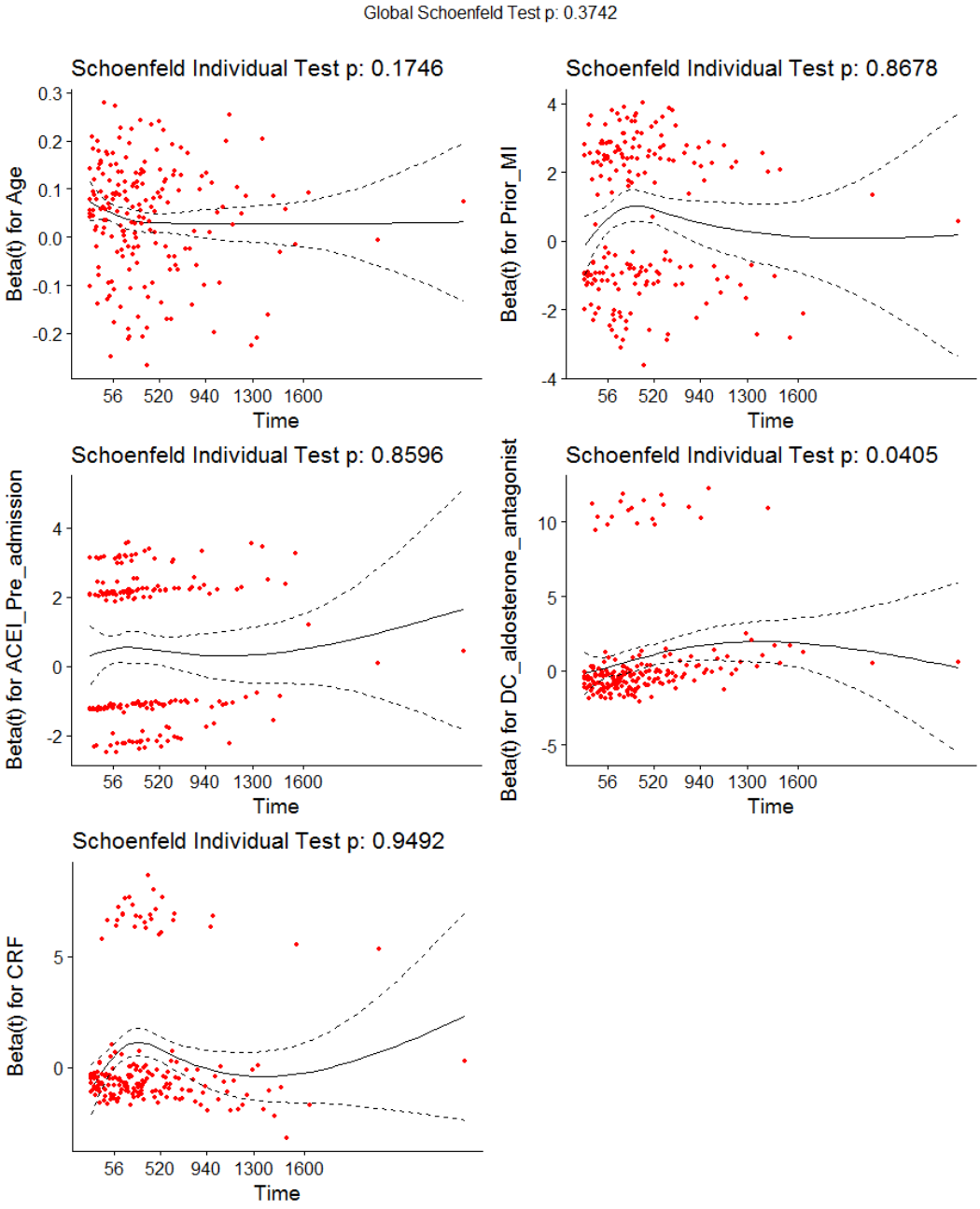


Figure 6.3: Schoenfeld residual plot for each significant clinical factor with the primary outcome.

primary outcome adjusting for significant clinical covariates and the first two principal components yielded six loci at genome-wide significance $\left(5 \times 10^{-8}\right)$. Table 6.4 is a summary of significant SNPs identified from the GWAS analysis. All significant SNPs have been investigated further individually through Kaplan-Meier plots comparing event survival by genotype, to distinguish the differences between carriers and non-carriers in the study cohort. LocusZoom plots were also produced to identify the LD between the significant SNPs and nearby variants based on European ancestry

individuals from the 1000 Genomes Project.

SNP	Gene	CHR	MAF	Info	HR	95% LCI	95% UCI	*p*-value
rs113348424	Unknown	14	0.013	0.706	8.806	4.283	18.107	3.33×10^{-9}
rs148409050	*IMMP2L*	7	0.011	0.883	6.346	3.397	11.854	6.80×10^{-9}
rs144599889	*IMMP2L*	7	0.011	0.882	6.305	3.382	11.756	6.90×10^{-9}
rs56045815	*CTNNA2*	2	0.030	0.906	3.102	2.082	4.621	2.60×10^{-8}
rs71472467	*INO80*	15	0.012	0.897	4.857	2.779	8.487	2.86×10^{-8}
rs34610018	Unknown	15	0.016	0.914	4.327	2.542	7.367	6.82×10^{-8}

Table 6.4: Summary of significant SNPs from single-variant primary outcome analysis. Abbreviations: CHR, chromosome; MAF, minor allele frequency; HR, hazard ratio; Info, info-score of imputation quality; LCI, lower confidence interval; UCI, upper confidence interval. *p*- value is calculated using the Wald test.

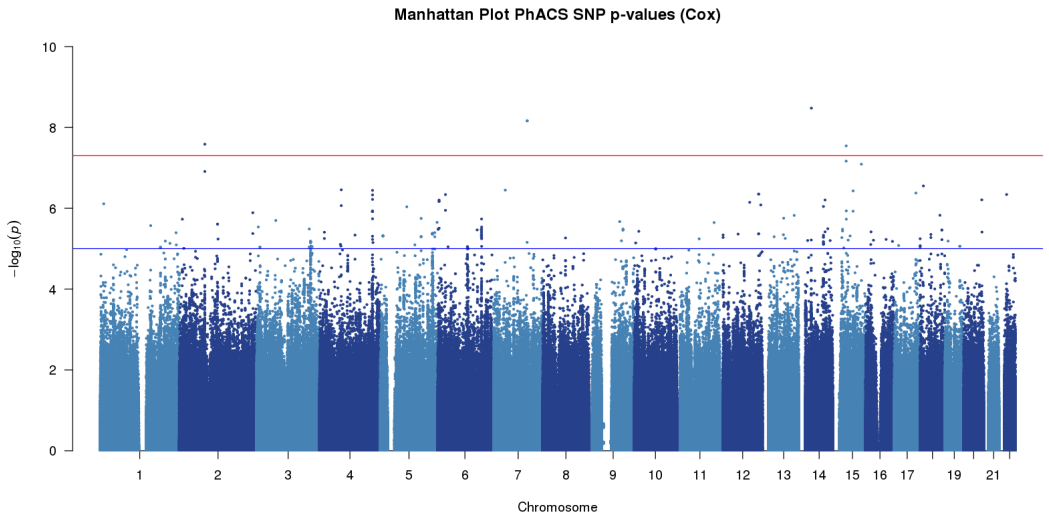


Figure 6.4: Manhattan plot results for the single variant Cox PH analysis of the primary outcome. Red line represents genome-wide significance threshold 5×10^{-8}. Blue line represents suggestive significance line. Each point represents a SNP.

The most significantly associated SNP with time to cardiovascular event was rs113348424 $\left(3.33 \times 10^{-9}\right)$ found on chromosome 14. The top left plot in Figure C.1 shows the cardiovascular event-free survival probability separated by genotype for the SNP rs113348424. From this, we can deduce that patients that carry the minor allele A are more likely to have a cardiovascular event relative to those that carry the G allele (HR = 8.806, 95% CI = 4.283-18.107). This SNP does not map to a gene but is located between the *LINC00639* and *SSTR1* genes. Figure B.1 shows that it is in very low LD ($r^2 < 0.4$) with two other SNPs.

The statistically significant SNPs rs148409050 and rs144599889 on chromosome 7 are located in the *IMMP2L* gene. Figure 6.6 shows that those with the heterozygous genotype TC at SNP rs148409050, have a significantly higher probability of a recurrent cardiovascular event occurring earlier than those with a homozygous genotype. A HR of 6.346 (95% CI = 3.397-11.854, $p = 6.80 \times 10^{-9}$), suggests that individuals carrying one copy of the C allele have an increased hazard of an event compared to individuals carrying two copies of the T allele. The median event-free survival time for patients with the heterozygous genotype is approximately 790 days compared to the T homozygous group which have a median event-free survival of approximately 1650 days. Likewise,

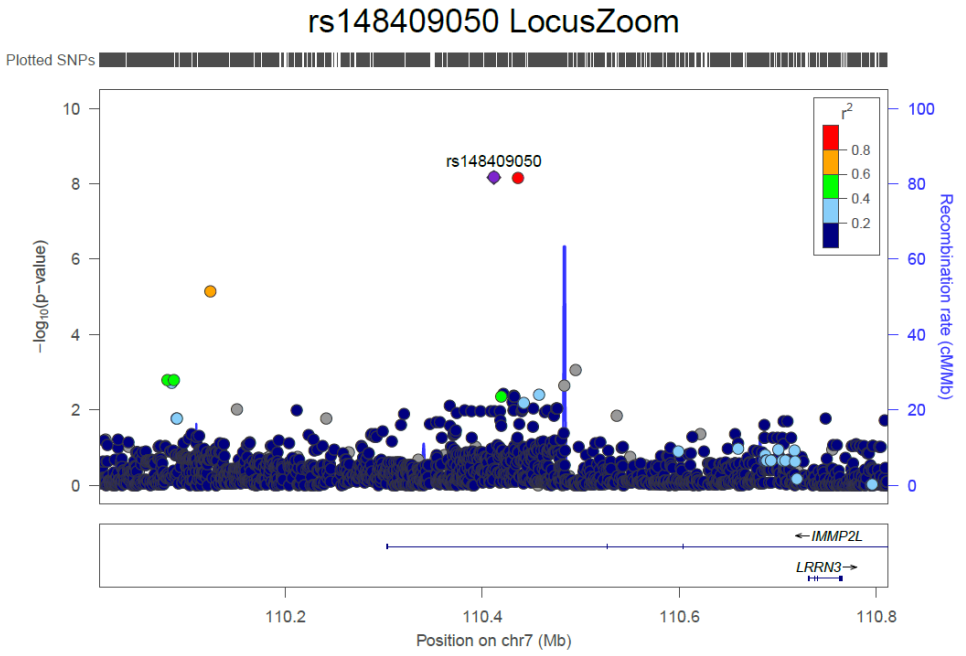


Figure 6.5: Association of rs148409050 with time to cardiovascular event. LocusZoom plot of the region associated with the primary outcome on chromosome 7 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs148409050 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs148409050 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

the top middle plot in Figure C.1, for rs144599889, shows that event-free survival is significantly reduced for those carrying the C allele in contrast to those that carry the T allele (HR = 6.305, 95% CI = 3.382-11.756, $p = 6.90 \times 10^{-9}$). The median survival for those with the heterozygous genotype is ≈780 days. Figure 6.5 shows that rs148409050 in the intron[2] region is highly correlated with its neighbouring SNP rs144599889 (also shown in Figure B.2). The SNP is also in moderate LD with many other variants within and near to the *IMMP2L* gene. *IMMP2L* is a protein-coding gene. This gene encodes a protein involved in processing the signal, peptide sequences used to direct mitochondrial proteins to the mitochondria (https://ghr.nlm.nih.gov/gene/IMMP2L#normalfunction). Mitochondria are organelles found in almost all cells. They essentially provide energy to the cell by converting oxygen and nutrients into adenosine triphosphate. Mitochondrial proteins are one of the necessary proteins for the catalytic activity of the mitochondrial

[2]An intron is the non-coding sequence in the gene. They can interrupt exons.

inner membrane peptidase complex.

The bottom left plot showing rs56045815 in Figure C.1 illustrates that the carriers of the C minor allele have a shorter survival time than those that have both copies of the G allele (HR = 3.102, 95% CI = 2.082-4.621, $p = 2.60 \times 10^{-8}$). The SNP rs56045815 is shown to be in moderate to high LD with another SNP just below genome-wide significance (see Figure B.3). rs56045815 (intron) is located in the protein-coding gene *CTNNA2*. Diseases associated with *CTNNA2* include mixed germ cell cancer. Among its related pathways are arrhythmogenic right ventricular cardiomyopathy and Sertoli-Sertoli cell junction dynamics (http://www.genecards.org/cgi-bin/carddisp.pl?gene=CTNNA2). With its association with right ventricular cardiomyopathy, *CTNNA2* may play a vital role in the likelihood of having a cardiovascular event. Right ventricular cardiomyopathy can weaken the walls of the ventricle by thinning and stretching the chambers, increasing the risk of sudden cardiac death.

The bottom middle plot in Figure C.1 shows that patients with one copy of the C allele at SNP rs71472467, have an increased hazard of a cardiovascular event than those with two copies of the A allele (HR = 4.857, 95% CI = 2.779-8.487, $p = 2.86 \times 10^{-8}$). rs71472467 (intron) on chromosome 15 located in the *INO80* gene, encodes a subunit of the chromatin remodelling complex. This protein is proposed to bind DNA and be recruited by the YY1 transcription factor to activate certain genes (https://www.ncbi.nlm.nih.gov/gene/54617). Figure B.4 shows that it is in high LD with a SNP in the *OIP5* gene and in moderate LD with the SNP, rs34610018, close to *INO80*, that is genome-wide significant for the primary outcome (see Figure B.5). The plot for rs34610018 shown in the top right corner of Figure C.1 shows that patients carrying one copy of the A allele have a median survival time of ≈1300 days compared to ≈1650 days in the group of individuals that carry two copies of the G allele. (HR = 4.327, 95% CI = 2.542-7.367, $p = 6.82 \times 10^{-8}$).

In summary, Figures 6.6 and C.1 all show that individuals with the heterozygous genotype for each significant SNP are at an increased risk of a cardiovascular event.

There is a large disparity between the sample size in each of the genotype groups. SNPs with a larger MAF would have a more balanced sample size within the groups contributing to increased power for detection.

Figure 6.7 shows multiple loci associated with time to all-cause mortality. A total of ten SNPs were found to be genome-wide significant $\left(5 \times 10^{-8}\right)$. The Cox PH model output for each of the SNPs is summarised in Table 6.5. The top left plot in Figure C.2 displays the survival curves for the SNP rs141689913. Patients who carry one copy of the allele A have significantly reduced survival compared to those that carry two copies of the T allele (HR = 6.424, 95% CI = 3.407-12.109, $p = 8.94 \times 10^{-9}$). The LocusZoom plot of SNP rs141689913 in Figure B.6 shows that the SNP is in low LD with five other neighbouring SNPs.

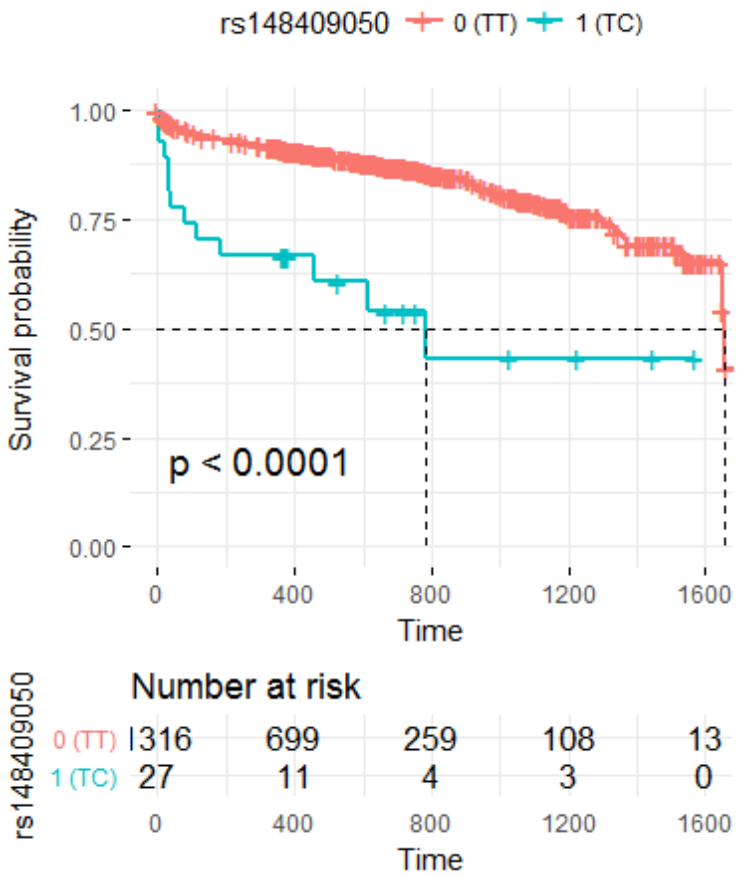


Figure 6.6: Kaplan-Meier plot by rs148409050 genotypes. T is the major allele and C is the minor allele. Time scale is in days. Median survival time is indicated with the black dotted line.

SNP	Gene	CHR	MAF	Info	HR	95% LCI	95% UCI	*p*-value
rs141689913	Unknown	7	0.015	0.663	6.424	3.407	12.109	8.94×10^{-9}
rs199571837	Unknown	14	0.035	0.739	4.593	2.703	7.803	1.73×10^{-8}
rs191847613	Unknown	14	0.029	0.805	4.568	2.674	7.806	2.72×10^{-8}
rs12402659	Unknown	1	0.028	0.926	4.463	2.633	7.564	2.76×10^{-8}
rs190226855	Unknown	12	0.021	0.895	4.606	2.680	7.914	3.21×10^{-8}
rs148484124	*URGCP - MRPS24*	7	0.017	0.902	4.869	2.750	8.622	5.60×10^{-8}
rs2695973	Unknown	5	0.028	0.964	0.212	0.121	0.371	5.76×10^{-8}
rs76428855	*MRPS25*	3	0.012	0.439	15.598	5.761	42.234	6.47×10^{-8}
rs141058803	Unknown	7	0.018	0.909	4.835	2.725	8.579	7.16×10^{-8}
rs141503732	*SSPO*	7	0.011	0.543	11.718	4.757	28.862	8.75×10^{-8}

Table 6.5: Summary of significant SNPs from single-variant secondary outcome analysis. Summary of significant SNPs from single-variant primary outcome analysis. Abbreviations: CHR, chromosome; MAF, minor allele frequency; HR, hazard ratio; Info, info-score of imputation quality; LCI, lower confidence interval; UCI, upper confidence interval. *p*- value is calculated using the Wald test.

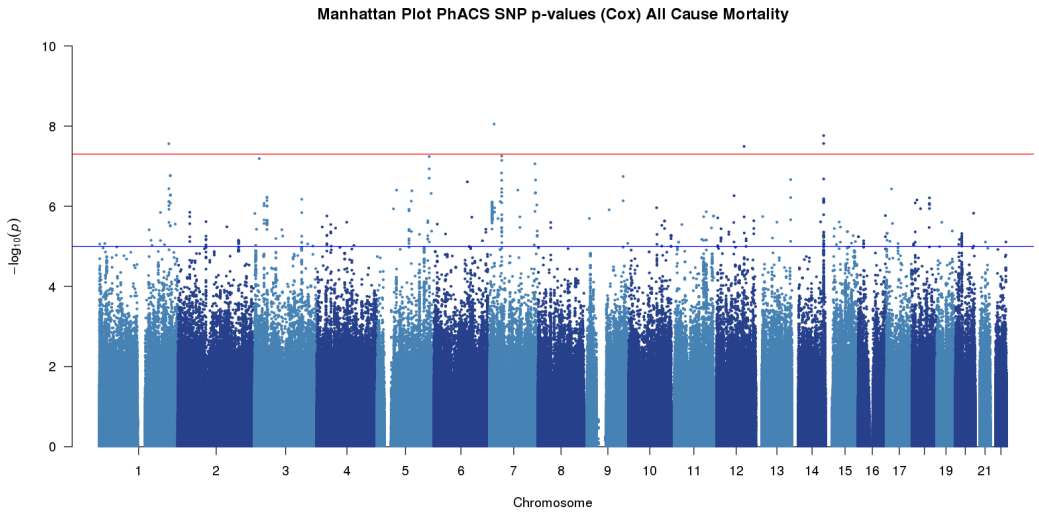


Figure 6.7: Manhattan plot results for the single variant Cox PH analysis of the secondary outcome. Red line represents genome-wide significance threshold 5×10^{-8}. Blue line represents suggestive significance line. Each point represents a SNP.

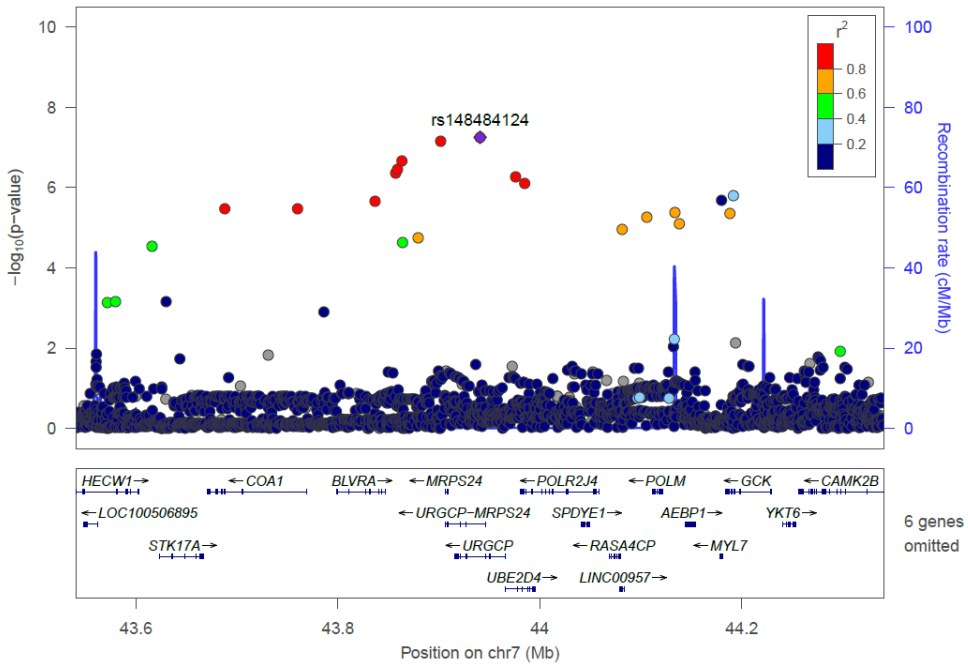


Figure 6.8: Association of rs148484124 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 7 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs148484124 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs148484124 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

According to the top middle and top right plots in Figure C.2, the SNPs rs199571837

(HR = 4.593, 95% CI = 2.703-7.803, $p = 1.73 \times 10^{-8}$) and rs191847613 (HR = 4.568, 95% CI = 2.674-7.806, $p = 2.72 \times 10^{-8}$) on chromosome 14 show that patients with one copy of the G allele have an increased hazard of mortality than those that carry both copies of the T allele. Figure B.7 and B.8 show the LocusZoom plots for rs199571837 and rs191847613, respectively. Both these SNPs map close to the *EML1* gene, with the plot of the locus indicating an extended haplotype of SNPs following the association of the lead SNP (rs199571837).

The Kaplan-Meier plot for SNP rs12402659 (middle left plot in Figure C.2) reports that patients carrying one copy of the T allele have a reduced survival probability than those carrying no copy of the T allele (HR = 4.463, 95% CI = 2.633-7.564, $p = 2.76 \times 10^{-8}$). Figure B.9 shows two SNPs in moderately high ($0.6 < r^2 < 0.8$) LD to the SNP rs12402659. One is located in the *MARK1* gene that is known to be associated with Alzheimer's disease (http://www.genecards.org/cgi-bin/carddisp.pl?gene=MARK1).

The central plot in Figure C.2 shows that carriers of the T allele at SNP rs190226855 have reduced survival compared to those that have no copies of the T allele (HR = 4.606, 95% CI = 2.680-7.914, $p = 3.21 \times 10^{-8}$). The regional plot for the SNP rs190226855, represented in Figure B.10 does not reveal any SNPs in high LD with the lead SNP.

The most interesting SNP from this discovery analysis was rs148484124 on chromosome 7 (Figure 6.8) because it is highly correlated with many neighbouring SNPs, including rs141058803 (Figure B.12) which is also associated at the genome-wide threshold with the secondary TTE outcome. There is also an abundance of different genes occupying this locus. rs148484124 is located in the *URGCP - MRPS24* gene. This locus represents naturally occurring read-through transcription between the neighbouring *URGCP* (up-regulator of cell proliferation) and *MRPS24* (mitochondrial ribosomal protein S24) genes (http://www.genecards.org/cgi-bin/carddisp.pl?gene=URGCP-MRPS24). Two SNPs in high LD with rs148484124 and rs141058803 are located in the *COA1* gene. Other SNPs are found in the *BLVRA*, *UBE2D4* and

POLR2J4 genes. However, none have any known association with the cardiovascular system. The Kaplan-Meier plots for both the SNPs rs148484124 (Figure 6.9) and rs141058803 (Bottom middle plot in Figure C.2) indicate patients with both copies of the major allele A have greater survival probability than those that only have one copy.

The locus plot (Figure B.11) for rs2695973 located on chromosome 5 shows some SNPs in moderate to low LD with the lead SNP, however, it does not map to a known gene. The middle right Kaplan-Meier plot in Figure C.2 for SNP rs2695973 shows a reduced probability of survival for the heterozygous genotype group.

SNP rs76428855 shown in the bottom left of Figure C.2 shows that median survival for patients with one copy of the C allele is ≈560 days (HR = 15.598, 95% CI = 5.761-42.234, $p = 6.47 \times 10^{-8}$). rs76428855 (intron) was found to be located in the protein-coding gene, *MRPS25*. Among its related pathways are Mitochondrial translation and organelle biogenesis and maintenance. *MRPS25* helps in protein synthesis within the mitochondrion (http://www.genecards.org/cgi-bin/carddisp.pl?gene=MRPS25). The LocusZoom plot (see Figure B.12) shows that the SNP is not correlated with any other variant.

The last significant SNP to map to a gene was rs141503732. The Kaplan-Meier plot (Figure C.2) indicates that patients with the heterozygous genotype are more likely to have a mortality causing event than those with the major homozygous genotype (HR = 11.718, 95% CI = 4.757-28.862, $p = 8.75 \times 10^{-8}$). This SNP is located in the *SSPO* (SCO-Spondin) gene, which is a protein-coding gene. Among its related pathways are HIV life cycle and O-linked glycosylation (http://www.genecards.org/cgi-bin/carddisp.pl?gene=SSPO). In summary, all the associated SNPs have a low MAF and would benefit from a replication study with larger sample size. The *URGCP - MRPS24* gene was the most promising candidate associated with all-cause mortality, containing the SNP rs148484124, which is correlated with a plethora of other associated SNPs in the flanking chromosome 7 region.

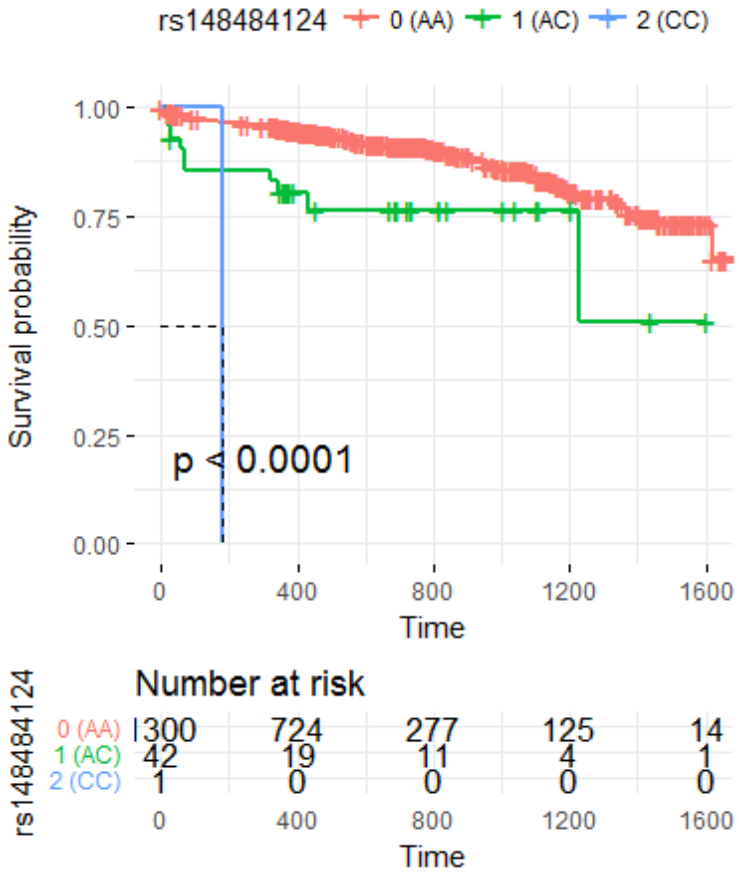


Figure 6.9: Kaplan-Meier plot by rs148484124 genotypes. A is the major allele and C is the minor allele. The timescale is in days. Median survival time is indicated with the black dotted line.

6.3.6 Rare Variant Association Analysis

Results for all 70663 transcripts included in the gene-based analysis of rare variants with the primary outcome are presented in Figure 6.10. The analysis using the BT with unit-weighting within a Cox PH model produced a total of 10 gene-outcome association signals found to be significant at a Bonferroni corrected threshold of 7.1×10^{-7} for the number of tests performed.

The top ranking gene that produced the lowest p-value was *PRKAG3* ($p = 8.00 \times 10^{-8}$) on chromosome 2. The gene contained 16 rare variants which, when collectively analysed using a BT within a Cox PH model produced a HR of 1.020 (95% CI = 1.013-1.028). This HR suggests that there is a very low increase in the hazard of a recurrent cardiovascular event for individuals with an increased burden of rare variants.

Gene	CHR	BP range	Variant Count	Total MAF	HR	95% LCI	95% UCI	p-value
PRKAG3	2	219,687,105-219,696,512	16	0.033	1.020	1.013	1.028	8.00×10^{-8}
ARRDC1	9	140,500,095-140,509,811	8	0.025	1.071	1.045	1.097	4.01×10^{-8}
LZTS2	10	102,756,964-102,767,585	16	0.041	1.038	1.024	1.052	4.30×10^{-8}
ADGRE1	19	6,887,581-6,940,463	136	0.341	1.007	1.005	1.010	2.17×10^{-9}
PABPC1L	20	43,538,702-43,567,962	68	0.146	1.008	1.006	1.012	1.30×10^{-8}
PHF13	1	6,673,755-6,684,092	11	0.015	1.029	1.017	1.040	6.89×10^{-7}
THAP3	1	6,685,209-6,695,645	28	0.064	1.013	1.008	1.018	3.56×10^{-7}
DNAJC11	1	6,694,227-6,761,966	115	0.250	1.004	1.003	1.006	4.80×10^{-7}
ZNF266	19	9,434,982-9,451,860	50	0.131	1.005	1.003	1.007	2.89×10^{-7}
SPINT4	20	44,350,987-44,354,335	14	0.024	1.029	1.018	1.040	3.17×10^{-7}

Table 6.6: Summary of significant genes from primary outcome analysis. Variant count includes all types of variants within the functional unit. Abbreviations: CHR, chromosome; BP, base pair; MAF, minor allele frequency; HR, hazard ratio; LCI, lower confidence interval; UCI, upper confidence interval. p- value is calculated using the Wald test.

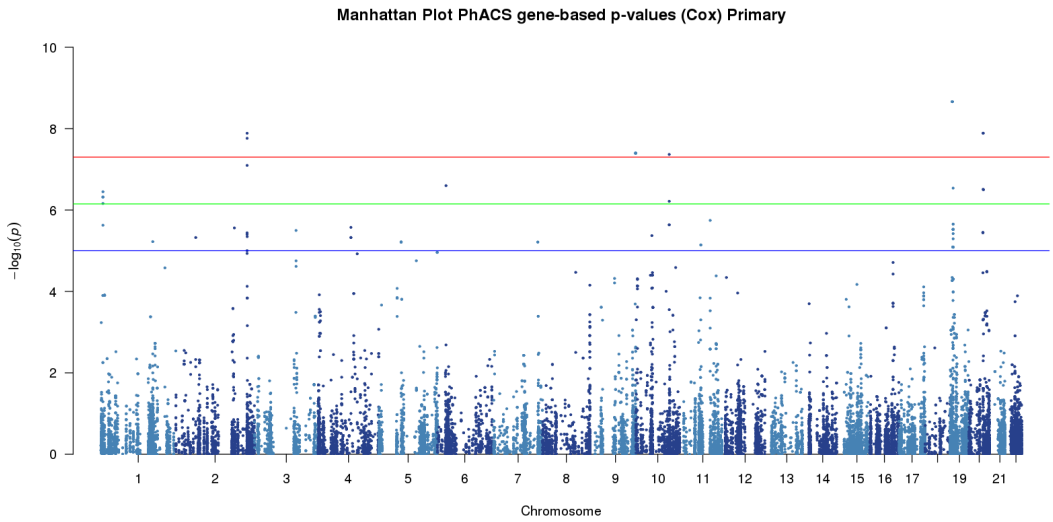


Figure 6.10: Manhattan plot results for the rare variant analysis of the primary outcome, using a BT within a Cox PH model. Red line represents genome-wide significance threshold 5×10^{-8}. The blue line represents suggestive significance line. Each point represents a gene.

The protein encoded by this gene is a regulatory subunit of the AMP-activated protein kinase. Genetic variants can be associated with increased glycogen content in skeletal muscle. The gene is expressed strongly in skeletal muscle, but more weakly in the heart and pancreas (https://www.ncbi.nlm.nih.gov/gene/53632).

The largest HR was for the protein-coding gene *ARRDC1*. This HR indicates that there is a 7.1% increase in hazard per unit increase in the (weighted) count of the 8 rare variants. *ARRDC1* plays a role in the extracellular transport of proteins between cells through the release in the extracellular space of microvesicles (https://www.uniprot.org/uniprot/Q8N5I2).

On chromosome 1, three genes found in close proximity to one another are all significantly associated with the primary outcome. *PHF13*, *THAP3* and *DNAJC11* are all protein-coding regions. *PHF13* modulates chromatin structure and has previously been associated with ovarian cancer survival time (http://www.genecards.org/cgi-bin/carddisp.pl?gene=PHF13). *THAP3* is highly expressed in the heart, skeletal muscle and placenta (http://www.uniprot.org/uniprot/Q8WTV1).

The remaining associated variants were not found to be directly biologically linked with the cardiovascular system through examination of literature and website searches. *LZTS2* is a protein coding gene that encodes the leucine zipper tumor suppressor family of proteins, which function in transcription regulation and cell cycle control. It is implicated in cancer, where it may inhibit cell proliferation and decrease susceptibility to tumor development (http://www.genecards.org/cgi-bin/carddisp.pl?gene=LZTS2). *ADGRE1* encodes a protein that has a domain resembling seven transmembrane G protein-coupled hormone receptors at its C-terminus (http://www.genecards.org/cgi-bin/carddisp.pl?gene=ADGRE1). *PABPC1L* is a protein coding gene. Diseases associated with *PABPC1L* include muscular dystrophy and rigid spine (http://www.uniprot.org/uniprot/Q4VXU2). *ZNF266* is a gene that encodes a protein containing many tandem zinc-finger motifs. Zinc fingers are protein or nucleic acid-binding domains, and may be involved in a variety of functions, including regulation of transcription. *ZNF266* is located in a cluster of similar genes encoding zinc finger proteins on chromosome 19 (http://www.genecards.org/cgi-bin/carddisp.pl?gene=ZNF266). *SPINT4* is a protein coding gene. GO annotations (http://geneontology.org/page/go-annotations) related to this gene include serine-type endopeptidase inhibitor activity (http://www.genecards.org/cgi-bin/carddisp.pl?gene=SPINT4).

In summary, according to the hazard ratios estimated in Table 6.6 for all significant genes, all the HRs above 1 indicate a reduction in median survival time for patients carrying minor alleles at rare variants, compared to those that do not.

The Manhattan plot for the gene-based analysis of the secondary outcome (see Figure 6.11) shows five genes to be significantly associated at a Bonferroni corrected threshold of 7.1×10^{-7}. The BT within a Cox PH model output is summarised in Table 6.7.

Gene	CHR	BP range	Variant Count	Total MAF	HR	95% LCI	95% UCI	p-value
ISCA1	9	88,879,463-88,897,490	26	0.061	1.034	1.022	1.047	7.32×10^{-8}
LIN52	14	74,181,825-74,227,001	90	0.175	1.007	1.004	1.009	1.76×10^{-8}
RPTOR	17	78,518,624-78,831,924	536	1.644	1.001	1.001	1.002	4.07×10^{-7}
DENND1C	19	6,467,218-6,481,798	40	0.074	1.014	1.009	1.019	1.49×10^{-7}
MRPL39	21	26,957,969-26,979,801	38	0.055	1.005	1.003	1.007	4.08×10^{-7}

Table 6.7: Summary of significant genes from secondary outcome analysis. Variant count includes all types of variants within the functional unit. Abbreviations: CHR, chromosome; BP, base pair; MAF, minor allele frequency; HR, hazard ratio; LCI, lower confidence interval; UCI, upper confidence interval. p- value is calculated using the Wald test.

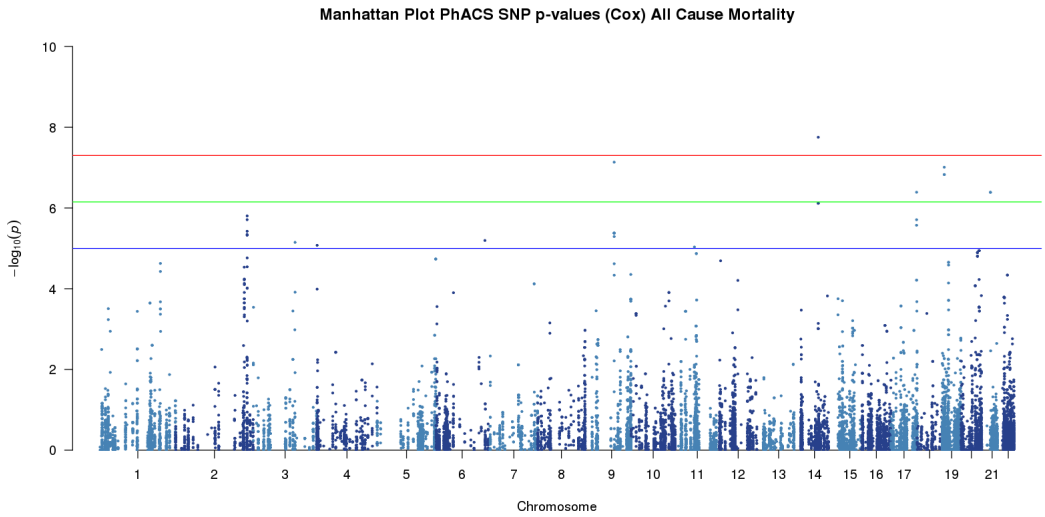


Figure 6.11: Manhattan plot results for the rare variant analysis of the secondary outcome, using a BT within a Cox PH model. Red line represents genome-wide significance threshold 5×10^{-8}. Blue line represents suggestive significance line. Each point represents a gene.

ISCA1 is the most significantly $\left(p = 7.32 \times 10^{-8}\right)$ associated gene with time to all-cause mortality. It is a mitochondrial protein-coding gene involved in the biogenesis and assembly of iron-sulfur clusters, which play a role in electron-transfer reactions. Gene expression has been detected in the cerebellum, kidney and heart (http://www.genecards.org/cgi-bin/carddisp.pl?gene=ISCA1).

MRPL39 is a protein-coding gene. Among its related pathways are mitochondrial translation and organelle biogenesis and maintenance. Two isoforms produced by alternative splicing, whereby Isoform 2 has heart-specific gene expression (http://www.genecards.org/cgi-bin/carddisp.pl?gene=MRPL39).

A direct link in terms of pathways, expression and literature indicating cardiovascular disease association could not be identified from the remaining significant protein coding genes: *RPTOR*, *LIN52* and *DENND1C*.

6.4 Discussion

The use of SurvivalGWAS_SV and rareSurvival for the analysis of the PhACS data successfully demonstrated the ability of the two programs to identify common and rare variant associations with TTE outcomes. The approximated runtime for the single-variant analysis of 8,788,380 SNPs of both the primary and secondary outcome using SurvivalGWAS_SV was a total of 59 hours. The computational runtime for the rare-variant analysis using rareSurvival was a slower process, amounting to a total of 572 hours which included both the primary and secondary outcome analyses. This was utilising the cluster specified in Section 6.2.

The analysis from the study yielded interesting results regarding the identification of novel loci and genes associated with time to recurrent cardiovascular events or mortality following an acute coronary event. The rare variant analysis using the BT within a Cox PH model identified candidate genes for further investigation. The programs addressed a need for the lack of software available to analyse genetic association studies with TTE outcomes. The PhACS study analysis demonstrated one of the first uses of gene-based testing of imputed genotype data using the BT within a Cox PH model for TTE outcomes.

The most interesting gene identified from the analyses of the primary outcome is *CTNNA2*, which has related pathways to right ventricular myopathy and contained SNP rs56045815, which showed a clear difference in event-free survival between genotype groups. *PRKAG3*, *PHF13*, *THAP3*, *DNAJC11* are all expressed in the heart. Nonetheless, all significant SNPs, including those not mapping directly to a gene, show significant differences between carrier and non-carriers of genotypes for recurrent cardiovascular event survival. The most interesting gene identified from the analyses of the secondary outcome was *URGCP-MRPS24*, which contains a large number of strongly associated SNPs in high LD with one another. The genes *ISCA1* and *MRPL39* are also of interest as both genes are expressed in the heart. These results suggest that all significant genes may warrant targeted sequencing in larger samples to confirm the

existence of rare survival influencing variants.

These findings demonstrate the effectiveness of aggregated tests of association in the identification of genes with time to cardiovascular event and mortality. For all associations, carrying minor alleles increased the risk of cardiovascular and mortality events. The statistical power of the association analyses is low because of the small sample size. Furthermore, caution should be taken when interpreting the significantly associated SNPs from the single variant analyses because the results are potentially unreliable due to the low MAFs of variants. Therefore any results would need to be replicated, and a computationally intensive follow-up test is recommended to distinguish true from false positives. Further analyses not reliant on asymptotic methods should be performed, such as permutation-based methods described by Wang et al. (2010). Other analyses can be undertaken on the raw PhACS study data and by using the GWAS results in this thesis. Details of the potential future perspective of PhACS is detailed in the final chapter.

CHAPTER 7

DISCUSSION AND CONCLUSION

7.1 Overview

This thesis details research into the evaluation and development of statistical methodology and computational tools to analyse genome-wide association studies (GWAS) of both common and rare genetic variants with time-to-event (TTE) outcomes.

SurvivalGWAS_Power, SurvivalGWAS_SV and rareSurvival will aid in the design of studies and identification of genetic biomarkers of patient response to treatment, with the ultimate goal of personalising therapeutic intervention for an array of diseases. Applying these computational tools to GWAS data has the potential for gene discovery to guide treatment choices, allowing the benefit/risk ratio to be optimised to achieve more prolonged survival in patients with diseases.

The preceding chapters have explained in detail many of the different aspects of this research. This discussion and conclusions chapter summarises the main findings from all previous chapters, highlighting the impact while calling attention to any limitations of the research. Recommendations are made to provide researchers with a basis for building on the current research of this thesis.

7.2 Implications of Research

Identifying genetic variants of treatment response has far-reaching implications in the field of precision medicine. The potential to personalise medicine for an individual will mark the end of general-purpose treatments for the entire population. This transition will increase the efficacy and safety (reduce adverse drug reactions) of treatments. To achieve this goal, statistical solutions play a key role in the study design and analytic phases of a study. Specifically, in the context of TTE studies, there is a shortage of

these computational tools to implement these statistical solutions.

Survival models, most commonly the Cox proportional hazards (PH) model, are being applied to genetic association studies of TTE outcomes. Even so, many studies opt to simplify the outcome to a binary response using models, such as logistic regression, in an attempt to lighten the computational burden of the analysis. Chapter 2 of this thesis demonstrated a clear advantage of using the Cox PH model of TTE outcomes over the logistic regression model of a simplified binary outcome through a simulation study across a wide variety of pharmacogenetic study designs. The comparison of methodology solidified the assertion that the Cox PH model would have greater power to detect associations for TTE outcomes than dichotomisation of the outcome. This research will impact the decision made by analytical teams on models used for the analysis of TTE outcomes. However even though this was commonly known in other research areas, the specific application to pharmacogenetics had not been previously considered.

The simulation studies and literature reviews throughout Chapters 2 to 5 highlighted the need for computational tools that can perform power calculations and analysis of both common and rare variants with TTE outcomes. The creation of bespoke software for this task has great advantages over general use programs such as R because they have been designed for those with limited computational programming knowledge and the ability to handle the scale and complexity of genetic data in pharmacogenetic GWAS.

This thesis presents the first software to simulate pharmacogenetic TTE data and performs power calculations based on two analytical models, the Cox proportional hazard and Weibull regression models (SurvivalGWAS_Power). The graphical user interface (GUI) provides a user-friendly program for Windows desktop users. SurvivalGWAS_Power will aid researchers in the design of pharmacogenetic TTE studies, providing details on the choice of statistical model and optimal sample size. This software was followed by two novel analytical tools: one for single variants, SurvivalGWAS_SV and another for gene-based tests of rare variant association, rareSurvival. Both tools

offer users simple command line execution of commands and compatibility with high-performance computing (HPC) clusters, resulting in efficient analyses in a field where "big data" translates to hundreds of terabytes. Both analytical tools will enable seamless analysis of large-scale genetic data in the hope of identifying variants that can help predict treatment response or prognosis for a number of diseases and traits. This target was demonstrated in the analysis of the Pharmacogenetics of Acute Coronary Syndrome (PhACS) study in Chapter 6, which identified some genes and variants that may have a potential impact on the occurrence of cardiovascular events and mortality.

The analysis performed in Chapter 6 was a demonstration of the application of the novel methodology and software developed in Chapters 4 and 5 of this thesis to identify novel relationships between single-variant and gene-based biomarkers for time to cardiovascular event and all-cause mortality in the PhACS data. In summary, the single variant analysis of the primary outcome produced 6 SNPs above genome-wide significance. The most interesting SNP identified was rs56045815, which maps to the *CTNNA2* gene, that is known to impact on arrhythmogenic right ventricular cardiomyopathy. The single variant analysis of the secondary outcome identified 10 SNPs at genome-wide significance. Although biological evidence was not found to associate these SNPs with all-cause mortality, further investigation is needed. The gene-based analyses of the primary outcome yielded ten genes significant at 7.1×10^{-7}. A locus containing the genes *PHF13*, *THAP3* and *DNAJC11* offers candidates for further exploration regarding function and pinpointing the causal variants. This statement is also true for the gene-based analysis of the secondary outcome, which identified six genes, two of which (*ISCA1* and *MRPL39*) are expressed in the heart. This result suggests potential candidates for future studies that may lead to targets for improved prediction of cardiovascular event outcomes.

7.3 Limitations

The research in this thesis is limited to comparing and contrasting two models, the Cox PH and Weibull regression models. Even though these are two common approaches, TTE data can be more complex. This includes different types of censored observation, multiple causes of an event and repeated measures of an event. The assumption of right censoring is used throughout this thesis. As is known for TTE data, patients can be left, interval and right censored as well as truncated, but this cannot currently be accommodated within the software. Figure 7.1 illustrates information collected on four patients follow-up times. Five different types of TTE outcomes are generated in this example, which requires different analyses to account for each of them. The software imple-

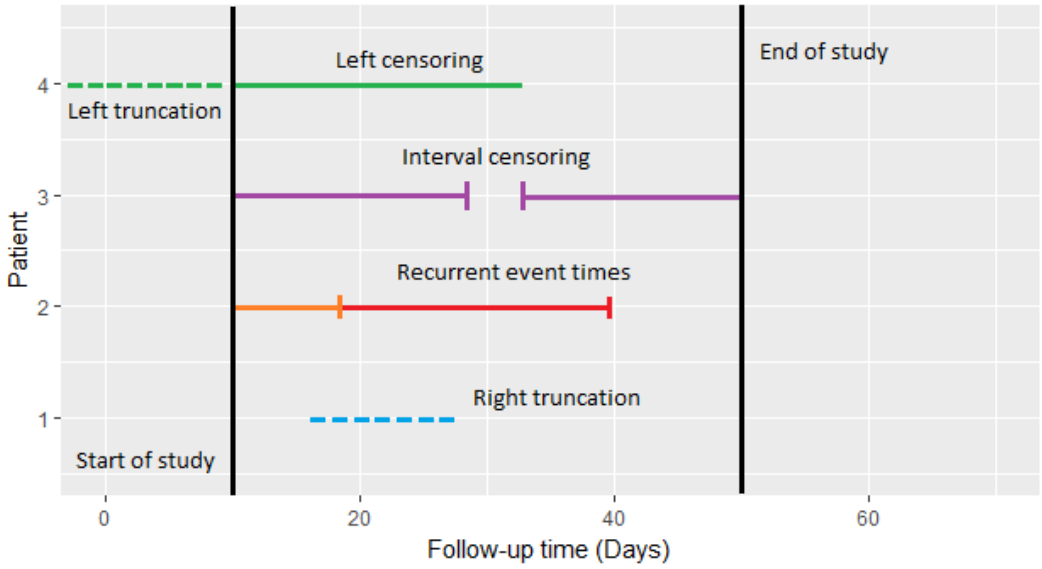


Figure 7.1: Follow-up times for four patients that experience different censoring, truncation and events.

menting the Weibull regression model is limited to the adjustment of a maximum of 10 additional non-genetic covariates. This model is further hindered by the convergence criteria used by the model to estimate coefficients. The Newton-Raphson algorithm implemented suffers from some inconsistency with estimating model parameters if the starting values are not close to the "true" value. Alternative convergence algorithms or extensions to the Newton-Raphson algorithm that account for the uncertainty of the

shape parameter in the Weibull model should be explored.

The use of C# to develop all three computational tools has been hindered by the lack of adaptability to Linux operating system computers and clusters, which are most commonly used within the field. Mono was used to compile the code for Linux and was the only option for this in 2014. In 2017, Microsoft Windows released Visual Studio Code and .NET Core which can produce fast running native Linux programs. Porting the code over from using the .NET framework to the .NET core is a short-term fix to gain improvement in software efficiency.

The findings of the study of the PhACS data should be considered with the caveat that the sample size was limited to 1367 patients, and there was no opportunity for replication of the findings. The rare variant analysis was undertaken with imputed data from the 1000 Genomes Project (Auton et al. 2015), rather than whole-exome or whole-genome sequencing, which limits the allele frequency to which reliable genotype prediction can be assured. Larger reference panels, such as the Haplotype Reference Consortium (The Haplotype Reference Consortium 2016), will enable high-quality imputation to lower allele frequencies, but sequencing remains the "gold standard".

7.4 Future Perspective

The development of methodology and computational tools for GWAS of common and rare variants with TTE outcomes has been undertaken in this thesis. There are many different aspects of this research that can be developed further, which includes novel methods and computational innovations that will enable more advanced application into the next generation of genetic data. This section outlines the short and long-term improvements, which can help develop SurvivalGWAS_Power, SurvivalGWAS_SV and rareSurvival, as well as the current state of survival analysis methodology.

Survival analysis methodology is evolving quickly within all fields of research, including genetics, with the majority of researchers implementing new methods within the R statistical environment. These methods can be used and adapted for application

within genetics research. Therefore, future versions of all three programs can employ more complex analysis techniques and extensions to account for more complex survival models. The Cox PH model does not adequately account for this complexity, especially when considering pharmacogenetic outcomes. This is particularly relevant when testing for associations between genetic variants and treatment failure, where the event of interest may occur due to multiple reasons such as lack of treatment efficacy or experiencing an adverse drug reaction. Each reason for drug failure may have distinct genetic and non-genetic risk factors, and such differences can be accommodated by considering each distinct outcome in a competing risks model. Competing risks are analysed using the Fine-Gray model (Fine & Gray 1999), which keeps those individuals who have already experienced the non-primary event within the risk set at a given time. These models provide insight into questions such as, what proportion of patients experience the event from cause k by time t, or what are the factors affecting the hazard of the event from cause k?

Other extensions to consider to make all three programs a more complete TTE design and analysis package would be to add options for parametric models, joint modelling of TTE and longitudinal data (Sudell et al. 2016) and accounting for recurrent events that are very common in studies involving epilepsy (Myers & Mefford 2015).

Measuring follow-up after all individuals have been recruited is an important feature of TTE studies. However, in many cases, an individual has an event immediately after recruitment or is censored before the follow-up time. These individuals are recorded as having an event time of zero, and they are typically removed from the analysis or assigned an event time that is very close to zero. This was observed in the analysis of PhACS in Chapter 6, where two patients were excluded on this basis. These individuals can be very informative for survival. Therefore, a solution is needed to handle and include this information in the analysis.

A final observation from the research in this thesis is that very little work had been published regarding appropriate adjustment for gene-level interaction effects. Papers

discussing gene-environment or gene-gene interactions are common for quantitative traits (Aschard 2016, Ma et al. 2013). Testing for gene-treatment interactions correctly can have great implications for pharmacogenetics research. GWAS of common and rare variants in pharmacogenetics still has the potential for drug development with the application into predictive/personalised medicine. However, the large effect sizes that most expected (Cirulli & Goldstein 2010) have been modest at best (Auer & Lettre 2015), and methodology development for interaction effects could help unlock information about treatment response.

7.4.1 SurvivalGWAS_Power

Chapter 3 discussed the development of power calculation software, covering the foundation of pharmacogenetic study designs and analysis models. The research in this chapter can be extended by first considering the simulation and testing of more realistic and complex scenarios, such as multiple treatment dose levels and accounting for flexibility with the treatment or dose administration time. Along with adding more sophistication to the treatment covariate, greater flexibility on the inclusion of additional covariates would be beneficial in the design and analysis of a study.

It is difficult to achieve a complete dataset when undertaking a study. The missing rate of variables cannot be avoided in both the genotype and phenotype data collected for all individuals. Adding an option into SurvivalGWAS_Power for increasing or decreasing the missing rate within data would give users an added benefit in designing realistic studies.

SurvivalGWAS_Power currently accounts for right censoring only. TTE studies have a wider variety of censoring and truncation[1] options such as left censoring, interval censoring, left truncation and right truncation. To analyse these options, the likelihoods of the statistical regression models need to be adapted.

SurvivalGWAS_Power currently simulates a SNP based on an additive genetic model.

[1]Patients excluded due to inadequate follow-up.

Providing users with the option of simulating a SNP with different genetic models (recessive and dominant) is of benefit. In many studies supplementation of SNP data through imputation is a common practice where genotype data is in the form of dosages calculated from genotype probabilities. The info-score is a useful metric of imputation quality, therefore allowing the user to adjust the info-score of a SNP similar to the control they have over MAF would help in determining the power to detect associations of the different analytical approaches with a range of imputation quality SNPs. SurvivalGWAS_Power provides some useful output metrics for users such as the distribution of estimated hazard ratios amongst replicates. Additional results that will be informative at the study design stage would be the probability of the disease/event occurring for each genotype group or allowing the user to specify the proportion of individuals dependent on genotype that will have the event of interest.

A short-term recommendation for SurvivalGWAS_Power would be to combine the C# code with R using R.NET. This package is a .NET framework that will enable calling in methods from libraries and scripts developed in R. A long-term improvement for SurvivalGWAS_Power, would be to eliminate the need for users to download a specific version distribution every time there is an update. Instead, the program can be designed as an interactive web application using R-Shiny (Chang et al. 2017). The software would be an interactive application with a GUI with the ability to access R libraries.

Calculating sample size to achieve appropriate power is very disease-specific (Sham & Purcell 2014). Therefore, a more advanced development of the program could be to express sample size and power calculation based on pilot data or data from similar studies to the one the user would like to design. This can be achieved by linking the application to an internet or MySQL database, which gathers and stores information from research articles or by direct upload into the program. Generating examples based on previous literature, provides more information for the user, resulting in more accurate power calculation to help drive decisions.

When conducting the simulation study in Chapter 5, it became apparent that there is software available for simulating rare variant data for binary and quantitative traits (Li et al. 2012, Chung & Shih 2013). Not only have the programs not been updated in several years, but they also do not provide options for TTE settings and models. An implementation within SurvivalGWAS_Power for simulating data based on rare-variants and calculating power for gene-based analysis models would be the first TTE application to offer this. Simulation of rare variants could use whole-exome or whole genome sequence data and population demographic models (Gazave et al. 2013, Excoffier et al. 2013) for a more realistic generation of variant data. Power analysis will be more accurate with realistic simulated data. The program SEQPower developed by Wang, Li, Lyn Santos-Cortez, Peng & Leal (2014) should be used as a bedrock to rare variant power analysis implementation within SurvivalGWAS_Power.

The current release and any extensions to SurvivalGWAS_Power will preemptively allow investigators the opportunity to cater for a wide variety of challenges faced when designing pharmacogenetic GWAS.

7.4.2 SurvivalGWAS_SV and rareSurvival

Similar extensions regarding new methodology implementation into SurvivalGWAS_-Power can be adapted for SurvivalGWAS_SV and rareSurvival. Currently, only the additive genetic model is assumed, whereas most genetic association software offers a choice of genetic models. Kim et al. (2013) developed a MaxTest for genomic data with TTE traits, which identifies the genetic model of each candidate SNP through a gradient lasso prediction model. This method could be investigated further through implementation in SurvivalGWAS_SV. Specific to rareSurvival, many different methodologies could be explored. Developing methodology that incorporates gene-based dispersion tests of association such as the SKAT or C-alpha statistic (Clarke et al. 2013) within a Cox PH model.

Short term recommendations for SurvivlGWAS_SV and rareSurvival would be to

port the code from the .NET framework to .NET Core, which is compatible with Linux and is substantially faster at running tasks than Mono. An alternative route to increase efficiency is to build an Apache Spark application in conjunction with Mobius (https://github.com/Microsoft/Mobius), a C# API. Apache Spark will provide efficient cluster and single computer management for any .NET framework language.

Both SurvivalGWAS_SV and rareSurvival have a multi-threading system where a specified number of threads need to be spawned, and these threads run a different part of the analysis. This implementation can be flawed because a user can specify too many threads or too few threads, whereby the process is slowed down as the computer cannot distribute the resources correctly. A pipeline pattern can rectify this problem, taking the choice away from the user and automatically assigning resources to the task. The pipeline algorithm processes a sequence of input parameters and executes concurrent queues of parallel tasks. This means that the program will be reading the data, analysing and outputting concurrently without slowing down or waiting for resources to become available.

Software development has now seen a dramatic change in the field when handling "big data". Programming languages such as Scala and Python are more commonly used with Apache Spark (Zaharia et al. 2016) for fast data analysis. Pipelines such as Hail and SeqSpark (Zhang et al. 2017) are at the forefront of this movement. A long-term upgrade for SurvivalGWAS_SV and rareSurvival is to develop a data input to output interpretation pipeline for genetic association studies with TTE outcomes. This pipeline can be achieved using both Scala and Python, which have the added benefit of compatibility with R, utilising all the libraries available. The algorithm concept of both software packages can be kept in place, but re-coded in Python. This platform would cover not only the main association analysis via single and gene-based tests but also plot Kaplan-Meier curves for covariates and variants and model checking for the PH assumption through Schoenfeld residuals. Ideally, conditional analyses after gene-based analyses would be implemented to localise the causal or multiple causal variant

associations. An Apache spark pipeline analysis tool will offer automated separation of data and a one-line command submission to consolidate multiple analyses into one program.

Rare variant association testing has mostly been at the gene-level, although some regions of the genome contain a smaller number of variants making the accumulation of the variants in these regions more vulnerable to being underpowered (many non-causal variants included in the pooling), and with a large amount of bias. Firth's test (Firth 1993) may provide a solution for rare variant tests, especially when the sample size is small and with a benefit over testing methods that rely on asymptotic assumptions that produce inflated type-I error rates (Wang 2014). The Firth procedure modifies the score function for the information matrix after maximisation of the Cox PH model, producing estimates by penalised maximum likelihood estimation.

After identifying the genomic regions in each of Chapters 4 and 5 simulation study for which our causal variants were located, the next step would be to pinpoint which variants within the gene are leading this signal of association. This is a crucial next step, with a need for further method development similar to Lin (2016) with the application into rareSurvival. Gene-based tests can hinder this process as these tests do not provide information at the individual locus and are ill-equipped to identify causal variants (Jeng et al. 2016). To address this, Jeng et al. (2016) had suggested rare variant association analysis at the single-locus level, proposing an adaptive false-negative control procedure.

7.4.3 PhACS

As mentioned in Chapter 6, the PhACS study can be investigated further. First, the rare-variant analysis identified functional units of significance with our TTE outcomes. An investigation into the source of these gene-based signals is needed, specifically by pinpointing whether a single or multiple variants are driving the association by using conditional analyses, eliminating each variant in turn based on the difference in p-value.

Some functional units have more than 500 variants to decipher the signal from, such as *RPTOR* in Table 6.7, which can prove to be a difficult task.

After the identification of variants through GWAS, a collection of evidence based on gene function can be undertaken in more detail through a thorough research synthesis of databases and previous literature. Conducting gene-set enrichment analysis could be of interest, to identify pathways linked to early treatment response, which may provide additional insight into relevant underlying biological and molecular processes for these outcomes. Most importantly, to obtain more knowledge on the effects of genetic variation on time to cardiovascular events, will require: (i) improved genomic annotation to establish the impact of genetic variation; (ii) establishment of causal association through replication based on different groups; and (iii) functional studies to determine gene function and regulation.

An analysis stratifying by drug use would be informative on the effectiveness of each treatment, with the possibility of investigating potential treatment interaction effects. A larger application could be to test SNP-treatment interaction for each cardiovascular drug under an additive dosage model after adjusting for clinical risk factors, and the main effects of SNP and treatment. These additional analyses may uncover other patterns of cardiovascular event-free survival.

A follow-up study with a larger sample size would be beneficial because, on average, the cumulative MAF of many of the gene regions analysed was less than 0.05, and the estimated HRs for both the primary and secondary outcome did not deviate from a value of 1 indicating that the hazard is the same for each group. Another source that potentially hindered the power to detect associations was that the true genetic architecture of cardiovascular event survival or mortality was unknown, therefore, the burden test within a Cox PH model may not have been the most powerful test for this scenario. This follows on from the earlier point made in this discussion chapter for using many different gene-based tests such as the SKAT on the data once implemented in future versions of rareSurvival.

7.5 Concluding Remarks

The work within this thesis has provided a detailed examination of GWAS with TTE outcomes. Insight through simulations comparing alternative regression approaches has solidified evidence of model choice under pharmacogenetic study designs with TTE outcomes. Novel software has been developed and tested with details provided on the future perspective of each computational tool. SurvivalGWAS_Power is important as it is the first genetic data simulator for TTE outcomes, and the first to enable estimation of power for multiple pharmacogenetic designs and analysis methods. SurvivalGWAS_SV provides users with an easy to use command line application, offering efficient analysis of large-scale genomic datasets using HPC clusters. SurvivalGWAS_SV implements two analytical approaches with an option to analyse SNP-covariate interaction effects. rareSurvival is the first rare-variant analysis tool for TTE outcomes, employing novel gene-based tests of association within TTE regression models. The use of the software in the analysis of the PhACS study data in Chapter 6 is an informative example of the potential use of the computational analysis tools and how beneficial they are to the research community. By making the software publicly available, it is envisaged that they will be applied to the analysis of whole-genome and -exome datasets with TTE outcomes in pharmacogenetics and other genetic studies to uncover the underlying mechanisms that affect complex human disease.

This research has implications within the study design phase of pharmacogenetic TTE studies through to the analysis phase. With the continued development of statistical methodology and computational tools for GWAS, we will understand more about the relationship between genetic variants and a multitude of phenotypes. Looking forward, GWAS will continue to expand the catalogue of loci of the genome contributing to complex human traits. The next phase of understanding the underlying biological mechanisms that cause disease will be through the collective knowledge of the genome, transcriptome (RNA), proteome, epigenome, metabolome and the methodological development to enable integration of these data resources. This will require co-ordinated

collaboration between researchers over a wide range of disciplines, including statistics, genetics, and computational biology. This, in turn, will contribute to the ultimate goal of many GWAS of complex diseases; the development of novel treatments and personalised medicine for an individual.

BIBLIOGRAPHY

Absenger, G., Benhaim, L., Szkandera, J., Zhang, W., Yang, D., Labonte, M. J., Pichler, M., Stotz, M., Samonigg, H., Renner, W., Gerger, A. & Lenz, H. J. (2014), 'The cyclin d1 (ccnd1) rs9344 g a polymorphism predicts clinical outcome in colon cancer patients treated with adjuvant 5-fu-based chemotherapy', *Pharmacogenomics J* **14**(2), 130–4.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23567490*

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Bonnen, P. E., de Bakker, P. I., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., Inouye, M., Jia, X., Palotie, A., Parkin, M., Whittaker, P., Chang, K., Hawes, A., Lewis, L. R., Ren, Y., Wheeler, D., Muzny, D. M., Barnes, C., Darvishi, K., Hurles, M., Korn, J. M., Kristiansson, K., Lee, C., McCarrol, S. A., Nemesh, J., Keinan, A., Montgomery, S. B., Pollack, S., Price, A. L., Soranzo, N., Gonzaga-Jauregui, C., Anttila, V., Brodeur, W., Daly, M. J., Leslie, S., McVean, G., Moutsianas, L., Nguyen, H., Zhang, Q., Ghori, M. J., McGinnis, R., McLaren, W., Takeuchi, F., Grossman, S. R., Shlyakhter, I., Hostetter, E. B., Sabeti, P. C., Adebamowo, C. A., Foster, M. W., Gordon, D. R., Licinio, J., Manca, M. C., Marshall, P. A., Matsuda, I., Ngare, D., Wang, V. O., Reddy, D., Rotimi, C. N., Royal, C. D., Sharp, R. R., Zeng, C., Brooks, L. D., McEwen, J. E. & Consortium, I. H. . (2010), 'Integrating common and rare genetic variation in diverse human populations', *Nature* **467**(7311), 52–8.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20811451*

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. & Zondervan, K. T. (2010), 'Data quality control in genetic case-control association studies', *Nat Protoc* **5**(9), 1564–73.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/21085122*

Aschard, H. (2016), 'A perspective on interaction effects in genetic association studies', *Genet Epidemiol* **40**(8), 678–688.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27390122*

Ashare, R. L., Karlawish, J. H., Wileyto, E. P., Pinto, A. & Lerman, C. (2013), 'Apoe ϵ4, an alzheimer's disease susceptibility allele, and smoking cessation', *Pharmacogenomics J* **13**(6), 538–43.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23247396*

Auer, P. L. & Lettre, G. (2015), 'Rare variant association studies: considerations, challenges and opportunities', *Genome Med* **7**(1), 16.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25709717*

Aulchenko, Y. S., Struchalin, M. V. & van Duijn, C. M. (2010), 'Probabel package for genome-wide association analysis of imputed data', *BMC Bioinformatics* **11**, 134.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20233392*

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., Abecasis, G. R. & Consortium, . G. P.

(2015), 'A global reference for human genetic variation', *Nature* **526**(7571), 68–74.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26432245*

BHF (2017), 'British heart foundation heart statistics'. Accessed: 2017-07-16.
URL: *https://www.bhf.org.uk/research/heart-statistics*

Bland, J. M. & Altman, D. G. (1995), 'Multiple significance tests: the bonferroni method', *BMJ* **310**(6973), 170.
URL: *https://www.bmj.com/content/310/6973/170*

Browning, B. L. & Browning, S. R. (2016), 'Genotype imputation with millions of reference samples', *Am J Hum Genet* **98**(1), 116–26.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26748515*

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. & Lee, J. J. (2015), 'Second-generation plink: rising to the challenge of larger and richer datasets', *GigaScience* **4**(1), 1–16.
URL: + *http://dx.doi.org/10.1186/s13742-015-0047-8*

Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. (2017), *shiny: Web Application Framework for R*. R package version 1.0.5.
URL: *https://CRAN.R-project.org/package=shiny*

Charland, S. L., Agatep, B. C., Herrera, V., Schrader, B., Frueh, F. W., Ryvkin, M., Shabbeer, J., Devlin, J. J., Superko, H. R. & Stanek, E. J. (2014), 'Providing patients with pharmacogenetic test results affects adherence to statin therapy: results of the additional kif6 risk offers better adherence to statins (akrobats) trial', *Pharmacogenomics J* **14**(3), 272–80.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23979174*

Chen, H., Lumley, T., Brody, J., Heard-Costa, N. L., Fox, C. S., Cupples, L. A. & Dupuis, J. (2014), 'Sequence kernel association test for survival traits', *Genet Epidemiol* **38**(3), 191–7.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24464521*

Chen, M. H. & Yang, Q. (2016), 'Rvfam: an r package for rare variant association analysis with family data', *Bioinformatics* **32**(4), 624–6.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26508760*

Chung, R. H. & Shih, C. C. (2013), 'Seqsimla: a sequence and phenotype simulation tool for complex disease studies', *BMC Bioinformatics* **14**, 199.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23782512*

Cirulli, E. T. & Goldstein, D. B. (2010), 'Uncovering the roles of rare variants in common disease through whole-genome sequencing', *Nat Rev Genet* **11**(6), 415–25.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20479773*

Clarke, G. M., Rivas, M. A. & Morris, A. P. (2013), 'A flexible approach for the analysis of rare variants allowing for a mixture of effects on binary or quantitative traits', *PLoS Genet* **9**(8), e1003694.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23966874*

Clarke, T. K., Crist, R. C., Ang, A., Ambrose-Lanci, L. M., Lohoff, F. W., Saxon, A. J., Ling, W., Hillhouse, M. P., Bruce, R. D., Woody, G. & Berrettini, W. H. (2014), 'Genetic variation in oprd1 and the response to treatment for opioid dependence with buprenorphine in european-american females', *Pharmacogenomics J* **14**(3), 303–8.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24126707*

Collett, D. (2003), *Modelling survival data in medical research*, third edition. edn, CRC Press.

Cox, D. R. (1975), 'Partial likelihood', *Biometrika* **62**(2), 269–276.
URL: + *http://dx.doi.org/10.1093/biomet/62.2.269*

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G. R. & Fuchsberger, C. (2016), 'Next-generation genotype imputation service and methods', *Nat Genet* **48**(10), 1284–1287.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27571263*

de With, S. A. J., Pulit, S. L., Staal, W. G., Kahn, R. S. & Ophoff, R. A. (2017), 'More than 25 years of genetic studies of clozapine-induced agranulocytosis', *Pharmacogenomics J* **17**(4), 304–311.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28418011*

Dean, L. (2015), Clopidogrel therapy and cyp2c19 genotype., *in* 'Medical Genetics Summaries [Internet].', Bethesda (MD): National Center for Biotechnology.
URL: *https://www.ncbi.nlm.nih.gov/books/NBK84114/*

Depta, J. P., Lenzini, P. A., Lanfear, D. E., Wang, T. Y., Spertus, J. A., Bach, R. G. & Cresci, S. (2015), 'Clinical outcomes associated with proton pump inhibitor use among clopidogrel-treated patients within cyp2c19 genotype groups following acute myocardial infarction', *Pharmacogenomics J* **15**(1), 20–5.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25001880*

Derkach, A., Lawless, J. F. & Sun, L. (2013), 'Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests', *Genet Epidemiol* **37**(1), 110–21.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23032573*

Devlin, B. & Roeder, K. (1999), 'Genomic control for association studies', *Biometrics* **55**(4), 997–1004.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.1999.00997.x*

Dudek, S. M., Motsinger, A. A., Velez, D. R., Williams, S. M. & Ritchie, M. D. (2006), 'Data simulation software for whole-genome association and other studies in human genetics', *Pac Symp Biocomput* pp. 499–510.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/17094264*

El Desoky, E. S., Derendorf, H. & Klotz, U. (2006), 'Variability in response to cardiovascular drugs', *Curr Clin Pharmacol* **1**(1), 35–46.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/18666376*

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. & Foll, M. (2013), 'Robust demographic inference from genomic and snp data', *PLoS Genet* **9**(10), e1003905.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24204310*

Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. (2016), 'The (in)famous gwas p-value threshold revisited and updated for low-frequency variants', *Eur J Hum Genet* **24**(8), 1202–5.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26733288*

Fernandez-Rozadilla, C., Cazier, J. B., Moreno, V., Crous-Bou, M., Guinó, E., Durán, G., Lamas, M. J., López, R., Candamio, S., Gallardo, E., Paré, L., Baiget, M., Páez, D., López-Fernández, L. A., Cortejoso, L., García, M. I., Bujanda, L., González, D., Gonzalo, V., Rodrigo, L., Reñé, J. M., Jover, R., Brea-Fernández, A., Andreu, M., Bessa, X., Llor, X., Xicola, R., Palles, C., Tomlinson, I., Castellví-Bel, S., Castells, A., Ruiz-Ponte, C., Carracedo, A. & Consortium, E. (2013), 'Pharmacogenomics in colorectal cancer: a genome-wide association study to predict toxicity after 5-fluorouracil or folfox administration', *Pharmacogenomics J* **13**(3), 209–17.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22310351*

Fine, J. P. & Gray, R. J. (1999), 'A proportional hazards model for the subdistribution of a competing risk', *Journal of the American Statistical Association* **94**(446), 496–509.
URL: *http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1999.10474144*

Firth, D. (1993), 'Bias reduction of maximum likelihood estimates', *Biometrika* **80**(1), 27–38.
URL: + *http://dx.doi.org/10.1093/biomet/80.1.27*

Franchini, M. (2016), 'Genetics of the acute coronary syndrome', *Ann Transl Med* **4**(10), 192.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27294088*

Gaastra, B., Shatunov, A., Pulit, S., Jones, A. R., Sproviero, W., Gillett, A., Chen, Z., Kirby, J., Fogh, I., Powell, J. F., Leigh, P. N., Morrison, K. E., Shaw, P. J., Shaw, C. E., van den Berg, L. H., Veldink, J. H., Lewis, C. M. & Al-Chalabi, A. (2016), 'Rare genetic variation in unc13a may modify survival in amyotrophic lateral sclerosis', *Amyotroph Lateral Scler Frontotemporal Degener* **17**(7-8), 593–599.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27584932*

Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., Boerwinkle, E., Gibbs, R. A., Sing, C. F., Clark, A. G. & Keinan, A. (2013), 'Neutral genomic regions refine models of recent rapid human population growth', *Proceedings of the National Academy of Sciences* .
URL: *http://www.pnas.org/content/early/2013/12/26/1310398110*

George, B., Seals, S. & Aban, I. (2014), 'Survival analysis and regression models', *J Nucl Cardiol* **21**(4), 686–94.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24810431*

Goldstein, J. I., Crenshaw, A., Carey, J., Grant, G. B., Maguire, J., Fromer, M., O'Dushlaine, C., Moran, J. L., Chambert, K., Stevens, C., , , Sklar, P., Hultman,

C. M., Purcell, S., McCarroll, S. A., Sullivan, P. F., Daly, M. J. & Neale, B. M. (2012), 'zcall: a rare variant caller for array-based genotypinggenetics and population analysis', *Bioinformatics* **28**(19), 2543–2545.
URL: + *http://dx.doi.org/10.1093/bioinformatics/bts479*

Gordon, A. S., Tabor, H. K., Johnson, A. D., Snively, B. M., Assimes, T. L., Auer, P. L., Ioannidis, J. P., Peters, U., Robinson, J. G., Sucheston, L. E., Wang, D., Sotoodehnia, N., Rotter, J. I., Psaty, B. M., Jackson, R. D., Herrington, D. M., O'Donnell, C. J., Reiner, A. P., Rich, S. S., Rieder, M. J., Bamshad, M. J., Nickerson, D. A. & Project, N. G. E. S. (2014), 'Quantifying rare, deleterious variation in 12 human cytochrome p450 drug-metabolism genes in a large-scale exome dataset', *Hum Mol Genet* **23**(8), 1957–63.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24282029*

Greenland, S., Pearl, J. & Robins, J. M. (1999), 'Causal diagrams for epidemiologic research', *Epidemiology* **10**(1), 37–48.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/9888278*

Gregers, J., Gréen, H., Christensen, I. J., Dalhoff, K., Schroeder, H., Carlsen, N., Rosthoej, S., Lausen, B., Schmiegelow, K. & Peterson, C. (2015), 'Polymorphisms in the abcb1 gene and effect on outcome and toxicity in childhood acute lymphoblastic leukemia', *Pharmacogenomics J* **15**(4), 372–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25582575*

Guo, Y., He, J., Zhao, S., Wu, H., Zhong, X., Sheng, Q., Samuels, D. C., Shyr, Y. & Long, J. (2014), 'Illumina human exome genotyping array clustering and quality control', *Nat Protoc* **9**(11), 2643–62.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25321409*

Han, J. Y., Lee, Y. S., Shin, E. S., Hwang, J. A., Nam, S., Hong, S. H., Ghang, H. Y., Kim, J. Y., Yoon, S. J. & Lee, J. S. (2014), 'A genome-wide association study of survival in small-cell lung cancer patients treated with irinotecan plus cisplatin chemotherapy', *Pharmacogenomics J* **14**(1), 20–7.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23478653*

He, L., Pitkäniemi, J., Heikkilä, K., Chou, Y. L., Madden, P. A., Korhonen, T., Sarin, A. P., Ripatti, S., Kaprio, J. & Loukola, A. (2016), 'Genome-wide time-to-event analysis on smoking progression stages in a family-based study', *Brain Behav* **6**(5), e00462.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27134767*

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. (2012), 'Fast and accurate genotype imputation in genome-wide association studies through pre-phasing', *Nat Genet* **44**(8), 955–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22820512*

Innocenti, F. (2005), *Pharmacogenomics : methods and applications*, Humana Press, Totowa, N.J.
URL: *http://www.springer.com/us/book/9781627034340*

Innocenti, F., Owzar, K., Cox, N. L., Evans, P., Kubo, M., Zembutsu, H., Jiang, C., Hollis, D., Mushiroda, T., Li, L., Friedman, P., Wang, L., Glubb, D., Hurwitz, H.,

Giacomini, K. M., McLeod, H. L., Goldberg, R. M., Schilsky, R. L., Kindler, H. L., Nakamura, Y. & Ratain, M. J. (2012), 'A genome-wide association study of overall survival in pancreatic cancer patients treated with gemcitabine in calgb 80303', *Clin Cancer Res* **18**(2), 577–84.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22142827*

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D. & Lin, X. (2013), 'Sequence kernel association tests for the combined effect of rare and common variants', *Am J Hum Genet* **92**(6), 841–53.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23684009*

Jeng, X. J., Daye, Z. J., Lu, W. & Tzeng, J. Y. (2016), 'Rare variants association analysis in large-scale sequencing studies at the single locus level', *PLoS Comput Biol* **12**(6), e1004993.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27355347*

Ji, Y., Biernacka, J. M., Hebbring, S., Chai, Y., Jenkins, G. D., Batzler, A., Snyder, K. A., Drews, M. S., Desta, Z., Flockhart, D., Mushiroda, T., Kubo, M., Nakamura, Y., Kamatani, N., Schaid, D., Weinshilboum, R. M. & Mrazek, D. A. (2013), 'Pharmacogenomics of selective serotonin reuptake inhibitor treatment for major depressive disorder: genome-wide associations and functional genomics', *Pharmacogenomics J* **13**(5), 456–63.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22907730*

Johnson, D. C., Weinhold, N., Mitchell, J. S., Chen, B., Kaiser, M., Begum, D. B., Hillengass, J., Bertsch, U., Gregory, W. A., Cairns, D., Jackson, G. H., Försti, A., Nickel, J., Hoffmann, P., Nöethen, M. M., Stephens, O. W., Barlogie, B., Davis, F. E., Hemminki, K., Goldschmidt, H., Houlston, R. S. & Morgan, G. J. (2016), 'Genome-wide association study identifies variation at 6q25.1 associated with survival in multiple myeloma', *Nat Commun* **7**, 10290.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26743840*

Johnson, J. L. (2017), 'Gas power calculator'.
URL: *http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/index.html*

Kamb, A., Harper, S. & Stefansson, K. (2013), 'Human genetics as a foundation for innovative drug development', *Nat Biotechnol* **31**(11), 975–8.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24213769*

Kanai, M., Tanaka, T. & Okada, Y. (2016), 'Empirical estimation of genome-wide significance thresholds based on the 1000 genomes project data set', *J Hum Genet* **61**(10), 861–866.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27305981*

Kapoor, M., Wang, J. C., Wetherill, L., Le, N., Bertelsen, S., Hinrichs, A. L., Budde, J., Agrawal, A., Almasy, L., Bucholz, K., Dick, D. M., Harari, O., Xiaoling, X., Hesselbrock, V., Kramer, J., Nurnberger, J. I., Rice, J., Schuckit, M., Tischfield, J., Porjesz, B., Edenberg, H. J., Bierut, L., Foroud, T. & Goate, A. (2014), 'Genome-wide survival analysis of age at onset of alcohol dependence in extended high-risk coga families', *Drug Alcohol Depend* **142**, 56–62.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24962325*

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. (2002), 'The human genome browser at ucsc', *Genome Research* **12**(6), 996–1006.
URL: *http://doi.org/10.1101/gr.229102*

Kim, J., Sohn, I., Son, D. S., Kim, D. H., Ahn, T. & Jung, S. H. (2013), 'Prediction of a time-to-event trait using genome wide snp data', *BMC Bioinformatics* **14**, 58.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23418752*

Koutras, A. K., Kotoula, V., Papadimitriou, C., Dionysopoulos, D., Zagouri, F., Kalofonos, H. P., Kourea, H. P., Skarlos, D. V., Samantas, E., Papadopoulou, K., Kosmidis, P., Pectasides, D. & Fountzilas, G. (2014), 'Vascular endothelial growth factor polymorphisms and clinical outcome in patients with metastatic breast cancer treated with weekly docetaxel', *Pharmacogenomics J* **14**(3), 248–55.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24061601*

Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. (2014), 'Rare-variant association analysis: study designs and statistical tests', *Am J Hum Genet* **95**(1), 5–23.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24995866*

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., Lin, X. & Team, N. G. E. S. P. L. P. (2012), 'Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies', *Am J Hum Genet* **91**(2), 224–37.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22863193*

Legge, S. E., Hamshere, M. L., Ripke, S., Pardinas, A. F., Goldstein, J. I., Rees, E., Richards, A. L., Leonenko, G., Jorskog, L. F., Chambert, K. D., Collier, D. A., Genovese, G., Giegling, I., Holmans, P., Jonasdottir, A., Kirov, G., McCarroll, S. A., MacCabe, J. H., Mantripragada, K., Moran, J. L., Neale, B. M., Stefansson, H., Rujescu, D., Daly, M. J., Sullivan, P. F., Owen, M. J., O'Donovan, M. C., Walters, J. T. R. & Consortium, C.-I. A. (2017), 'Genome-wide common and rare variant analysis provides novel insights into clozapine-associated neutropenia', *Mol Psychiatry* **22**(10), 1509.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27502474*

Lemieux Perreault, L. P., Legault, M. A., Asselin, G. & Dubé, M. P. (2016), 'genipe: an automated genome-wide imputation pipeline with automatic reporting and statistical tools', *Bioinformatics* **32**(23), 3661–3663.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27497439*

Leschziner, G., Jorgensen, A. L., Andrew, T., Pirmohamed, M., Williamson, P. R., Marson, A. G., Coffey, A. J., Middleditch, C., Rogers, J., Bentley, D. R., Chadwick, D. W., Balding, D. J. & Johnson, M. R. (2006), 'Clinical factors and abcb1 polymorphisms in prediction of antiepileptic drug response: a prospective cohort study', *Lancet Neurol* **5**(8), 668–76.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/16857572*

Li, B., Wang, G. & Leal, S. M. (2012), 'Simrare: a program to generate and analyze sequence-based data for association studies of quantitative and qualitative traits', *Bioinformatics* **28**(20), 2703–4.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22914216*

Li, H. (2011), 'A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data', *Bioinformatics* **27**(21), 2987–2993.

Lin, W. Y. (2016), 'Beyond rare-variant association testing: Pinpointing rare causal variants in case-control sequencing study', *Sci Rep* **6**, 21824.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26903168*

Lin, W. Y., Lou, X. Y., Gao, G. & Liu, N. (2014), 'Rare variant association testing by adaptive combination of p-values', *PLoS One* **9**(1), e85728.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24454922*

Lin, X., Cai, T., Wu, M. C., Zhou, Q., Liu, G. & Christiani, D. C. (2011), 'Kernel machine snp-set analysis for censored survival outcomes in genome-wide association studies', *Genet Epidemiol* **35**(7), 620–31.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/21818772*

Loh, P. R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N. & Price, A. L. (2015), 'Efficient bayesian mixed-model analysis increases association power in large cohorts', *Nat Genet* **47**(3), 284–90.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25642633*

Lohoff, F. W., Narasimhan, S. & Rickels, K. (2013), 'Interaction between polymorphisms in serotonin transporter (slc6a4) and serotonin receptor 2a (htr2a) genes predict treatment response to venlafaxine xr in generalized anxiety disorder', *Pharmacogenomics J* **13**(5), 464–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22907732*

Low, S. K., Takahashi, A., Mushiroda, T. & Kubo, M. (2014), 'Genome-wide association study: a useful tool to identify common genetic variants associated with drug toxicity and efficacy in cancer pharmacogenomics', *Clin Cancer Res* **20**(10), 2541–52.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24831277*

Ma, L., Clark, A. G. & Keinan, A. (2013), 'Gene-based testing of interactions in association studies of quantitative traits', *PLoS Genet* **9**(2), e1003321.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23468652*

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F. & Parkinson, H. (2017), 'The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog)', *Nucleic Acids Res* **45**(D1), D896–D901.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27899670*

Machiela, M. J. & Chanock, S. J. (2015), 'Ldlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants', *Bioinformatics* **31**(21), 3555–3557.
URL: + *http://dx.doi.org/10.1093/bioinformatics/btv402*

Mackelprang, R. D., Bamshad, M. J., Chong, J. X., Hou, X., Buckingham, K. J., Shively, K., deBruyn, G., Mugo, N. R., Mullins, J. I., McElrath, M. J., Baeten, J. M., Celum, C., Emond, M. J., Lingappa, J. R., Teams, P. i. P. H. T. S. & the Partners PrEP Study (2017), 'Whole genome sequencing of extreme phenotypes identifies variants in cd101 and ube2v1 associated with increased risk of sexually acquired hiv-1', *PLoS Pathog* **13**(11), e1006703.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/29108000*

Madsen, B. E. & Browning, S. R. (2009), 'A groupwise association test for rare mutations using a weighted sum statistic', *PLoS Genet* **5**(2), e1000384.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19214210*

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. & Visscher, P. M. (2009), 'Finding the missing heritability of complex diseases', *Nature* **461**(7265), 747–53.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19812666*

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. (2007), 'A new multipoint method for genome-wide association studies by imputation of genotypes', *Nat Genet* **39**(7), 906–13.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/17572673*

Morris, A. P. & Zeggini, E. (2010), 'An evaluation of statistical approaches to rare variant analysis in genetic association studies', *Genet Epidemiol* **34**(2), 188–93.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/19810025*

Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., Albers, P. K., McVean, G., Boehnke, M., Altshuler, D., McCarthy, M. I. & Consortium, G. (2015), 'The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease', *PLoS Genet* **11**(4), e1005165.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25906071*

Moutsianas, L. & Morris, A. P. (2014), 'Methodology for the analysis of rare genetic variation in genome-wide association and re-sequencing studies of complex human traits', *Brief Funct Genomics* **13**(5), 362–70.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24916163*

Myers, C. T. & Mefford, H. C. (2015), 'Advancing epilepsy genetics in the genomic era', *Genome Med* **7**, 91.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26302787*

Mägi, R., Asimit, J. L., Day-Williams, A. G., Zeggini, E. & Morris, A. P. (2012), 'Genome-wide association analysis of imputed rare variants: application to seven common complex diseases', *Genet Epidemiol* **36**(8), 785–96.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22951892*

Mägi, R., Kumar, A. & Morris, A. P. (2011), 'Assessing the impact of missing genotype data in rare variant association analysis', *BMC Proc* **5 Suppl 9**, S107.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22373025*

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., Indap, A., King, K. S., Bergmann, S., Nelson, M. R., Stephens, M. & Bustamante, C. D. (2008), 'Genes mirror geography within europe', *Nature* **456**(7218), 98–101.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/18758442*

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., Navarro, P., Zagury, J.-F., Wilson, J. F., Toniolo, D., Gasparini, P., Soranzo, N., Sandhu, M. S. & Marchini, J. (2014), 'A general approach for haplotype phasing across the full spectrum of relatedness', *PLOS Genetics* **10**(4), 1–21.
URL: *https://doi.org/10.1371/journal.pgen.1004234*

Owzar, K., Li, Z., Cox, N. & Jung, S. H. (2012), 'Power and sample size calculations for snp association studies with censored time-to-event outcomes', *Genet Epidemiol* **36**(6), 538–48.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/22685040*

Panagiotou, O. A., Ioannidis, J. P. A. & for the Genome-Wide Significance Project (2012), 'What should the genome-wide significance threshold be? empirical replication of borderline genetic associations', *International Journal of Epidemiology* **41**(1), 273–286.
URL: *http://dx.doi.org/10.1093/ije/dyr178*

Pander, J., van Huis-Tanja, L., Böhringer, S., van der Straaten, T., Gelderblom, H., Punt, C. & Guchelaar, H. J. (2015), 'Genome wide association study for predictors of progression free survival in patients on capecitabine, oxaliplatin, bevacizumab and cetuximab in first-line therapy of metastatic colorectal cancer', *PLoS One* **10**(7), e0131091.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26222057*

Pang, S. Y., Hsu, J. S., Teo, K. C., Li, Y., Kung, M. H. W., Cheah, K. S. E., Chan, D., Cheung, K. M. C., Li, M., Sham, P. C. & Ho, S. L. (2017), 'Burden of rare variants in als genes influences survival in familial and sporadic als', *Neurobiol Aging* **58**, 238.e9–238.e15.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28709720*

Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. (2008), 'Estimation of the multiple testing burden for genomewide association studies of nearly all common variants', *Genet Epidemiol* **32**(4), 381–5.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/18348202*

Phipps, A. I., Passarelli, M. N., Chan, A. T., Harrison, T. A., Jeon, J., Hutter, C. M., Berndt, S. I., Brenner, H., Caan, B. J., Campbell, P. T., Chang-Claude, J., Chanock, S. J., Cheadle, J. P., Curtis, K. R., Duggan, D., Fisher, D., Fuchs, C. S., Gala, M., Giovannucci, E. L., Hayes, R. B., Hoffmeister, M., Hsu, L., Jacobs, E. J., Jansen, L., Kaplan, R., Kap, E. J., Maughan, T. S., Potter, J. D., Schoen, R. E., Seminara, D.,

Slattery, M. L., West, H., White, E., Peters, U. & Newcomb, P. A. (2016), 'Common genetic variation and survival after colorectal cancer diagnosis: a genome-wide analysis', *Carcinogenesis* **37**(1), 87–95.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26586795*

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006), 'Principal components analysis corrects for stratification in genome-wide association studies', *Nat Genet* **38**(8), 904–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/16862161*

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R. & Willer, C. J. (2010), 'Locuszoom: regional visualization of genome-wide association scan results', *Bioinformatics* **26**(18), 2336–2337.
URL: + *http://dx.doi.org/10.1093/bioinformatics/btq419*

Purcell, S., Cherny, S. S. & Sham, P. C. (2003), 'Genetic power calculator: design of linkage and association genetic mapping studies of complex traits', *Bioinformatics* **19**(1), 149–50.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/12499305*

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J. & Sham, P. C. (2007), 'Plink: a tool set for whole-genome association and population-based linkage analyses', *Am J Hum Genet* **81**(3), 559–75.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/17701901*

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Ray, A., Tennakoon, L., Keller, J., Sarginson, J. E., Ryan, H. S., Murphy, G. M., Lazzeroni, L. C., Trivedi, M. H., Kocsis, J. H., DeBattista, C. & Schatzberg, A. F. (2015), 'Abcb1 (mdr1) predicts remission on p-gp substrates in chronic depression', *Pharmacogenomics J* **15**(4), 332–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25487678*

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A. & Lancet, D. (2010), 'Genecards version 3: the human gene integrator', *Database (Oxford)* **2010**, baq020.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20689021*

Santorico, S. A. & Hendricks, A. E. (2016), 'Progress in methods for rare variant association', *BMC Genet* **17 Suppl 2**, 6.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26866487*

Sato, Y., Yamamoto, N., Kunitoh, H., Ohe, Y., Minami, H., Laird, N. M., Katori, N., Saito, Y., Ohnami, S., Sakamoto, H., Sawada, J., Saijo, N., Yoshida, T. & Tamura, T. (2011), 'Genome-wide association study on overall survival of advanced non-small cell lung cancer patients treated with carboplatin and paclitaxel', *J Thorac Oncol* **6**(1), 132–8.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/21079520*

Schoenfeld, D. (1982), 'Partial residuals for the proportional hazards regression model', *Biometrika* **69**(1), 239–241.
URL: *http://www.jstor.org/stable/2335876*

Seed, C., Bloemendal, A., Bloom, J. M., Goldstein, J. I., King, D., Poterba, T. & Neale., B. M. (2017), 'Hail: An open-source framework for scalable genetic data analysis.'. In preparation.
URL: *https://github.com/hail-is/hail*

Sham, P. C. & Purcell, S. M. (2014), 'Statistical power and significance testing in large-scale genetic studies', *Nat Rev Genet* **15**(5), 335–46.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24739678*

Shin, J. & Johnson, J. A. (2010), 'Beta-blocker pharmacogenetics in heart failure', *Heart Fail Rev* **15**(3), 187–96.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/18437562*

Shmueli, G. (2010), 'To explain or to predict?', *Statist. Sci.* **25**(3), 289–310.
URL: *https://doi.org/10.1214/10-STS330*

Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. (2006), 'Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies', *Nat Genet* **38**(2), 209–13.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/16415888*

Souza, C. R. (2014), 'The accord.net framework'.
URL: *http://accord-framework.net*

Speed, D., Hoggart, C., Petrovski, S., Tachmazidou, I., Coffey, A., Jorgensen, A., Eleftherohorinou, H., De Iorio, M., Todaro, M., De, T., Smith, D., Smith, P. E., Jackson, M., Cooper, P., Kellett, M., Howell, S., Newton, M., Yerra, R., Tan, M., French, C., Reuber, M., Sills, G. E., Chadwick, D., Pirmohamed, M., Bentley, D., Scheffer, I., Berkovic, S., Balding, D., Palotie, A., Marson, A., O'Brien, T. J. & Johnson, M. R. (2014), 'A genome-wide association study and biological pathway analysis of epilepsy prognosis in a prospective cohort of newly treated epilepsy', *Hum Mol Genet* **23**(1), 247–58.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23962720*

Su, Z., Marchini, J. & Donnelly, P. (2011), 'Hapgen2: simulation of multiple disease snps', *Bioinformatics* **27**(16), 2304–5.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/21653516*

Subirana, I. & González, J. R. (2013), 'Genetic association analysis and meta-analysis of imputed snps in longitudinal studies', *Genet Epidemiol* **37**(5), 465–77.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23595425*

Sudell, M., Kolamunnage-Dona, R. & Tudur-Smith, C. (2016), 'Joint models for longitudinal and time-to-event data: a review of reporting quality with a view to meta-analysis', *BMC Medical Research Methodology* **16**(1), 168.
URL: *https://doi.org/10.1186/s12874-016-0272-6*

Teare, M. D. (2011), *Genetic epidemiology*, Humana Press, New York ; London.
URL: *1850-9999 http://www.springer.com/gb/ BLDSS*

The Haplotype Reference Consortium, . (2016), 'A reference panel of 64,976 haplotypes for genotype imputation', *Nature Genetics* **48**, 1279–1283.
URL: *http://dx.doi.org/10.1038/ng.3643*

Therneau, T. M. (2015), *A Package for Survival Analysis in S*. version 2.38.
URL: *https://CRAN.R-project.org/package=survival*

Therneau, T. M., Grambsch, P. M. & Fleming, T. R. (1990), 'Martingale-based residuals for survival models', *Biometrika* **77**(1), 147–160.
URL: + *http://dx.doi.org/10.1093/biomet/77.1.147*

Turner, R. M., Park, B. K. & Pirmohamed, M. (2015), 'Parsing interindividual drug variability: an emerging role for systems pharmacology', *Wiley Interdiscip Rev Syst Biol Med* **7**(4), 221–41.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25950758*

Turner, R. M., Yin, P., Hanson, A., FitzGerald, R., Morris, A. P., Stables, R. H., Jorgensen, A. L. & Pirmohamed, M. (2017), 'Investigating the prevalence, predictors, and prognosis of suboptimal statin use early after a non-st elevation acute coronary syndrome', *J Clin Lipidol* **11**(1), 204–214.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/28391887*

UniProt (2017), 'Uniprot: the universal protein knowledgebase', *Nucleic Acids Research* **45**(D1), D158–D169.
URL: *http://dx.doi.org/10.1093/nar/gkw1099*

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M. & Wei, L. J. (2014), 'Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis', *J Clin Oncol* **32**(22), 2380–5.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24982461*

Uppugunduri, C. R., Rezgui, M. A., Diaz, P. H., Tyagi, A. K., Rousseau, J., Daali, Y., Duval, M., Bittencourt, H., Krajinovic, M. & Ansari, M. (2014), 'The association of cytochrome p450 genetic polymorphisms with sulfolane formation and the efficacy of a busulfan-based conditioning regimen in pediatric patients undergoing hematopoietic stem cell transplantation', *Pharmacogenomics J* **14**(3), 263–71.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24165757*

Vandin, F., Papoutsaki, A., Raphael, B. J. & Upfal, E. (2015), 'Accurate computation of survival statistics in genome-wide studies', *PLoS Comput Biol* **11**(5), e1004071.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25950620*

Verma, S. S., de Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou-Khales, B., Mukherjee, S., Jarvik, G. P., Kottyan, L. C., Burt, A., Bradford, Y., Armstrong, G. D., Derr, K., Crawford, D. C., Haines, J. L., Li, R., Crosslin, D. & Ritchie, M. D. (2014), 'Imputation and quality control steps for combining multiple genome-wide datasets', *Front Genet* **5**, 370.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25566314*

Voorman, A., Brody, J., Chen, H., Lumley, T. & Davis, B. (2017), *seqMeta: Meta-Analysis of Region-Based Tests of Rare DNA Variants*. R package version 1.6.7.
URL: *https://CRAN.R-project.org/package=seqMeta*

Wagner, M. J. (2013), 'Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits', *Pharmacogenomics* **14**(4), 413–24.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23438888*

Walter, S., Atzmon, G., Demerath, E. W., Garcia, M. E., Kaplan, R. C., Kumari, M., Lunetta, K. L., Milaneschi, Y., Tanaka, T., Tranah, G. J., Völker, U., Yu, L., Arnold, A., Benjamin, E. J., Biffar, R., Buchman, A. S., Boerwinkle, E., Couper, D., De Jager, P. L., Evans, D. A., Harris, T. B., Hoffmann, W., Hofman, A., Karasik, D., Kiel, D. P., Kocher, T., Kuningas, M., Launer, L. J., Lohman, K. K., Lutsey, P. L., Mackenbach, J., Marciante, K., Psaty, B. M., Reiman, E. M., Rotter, J. I., Seshadri, S., Shardell, M. D., Smith, A. V., van Duijn, C., Walston, J., Zillikens, M. C., Bandinelli, S., Baumeister, S. E., Bennett, D. A., Ferrucci, L., Gudnason, V., Kivimaki, M., Liu, Y., Murabito, J. M., Newman, A. B., Tiemeier, H. & Franceschini, N. (2011), 'A genome-wide association study of aging', *Neurobiol Aging* **32**(11), 2109.e15–28.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/21782286*

Wang, G. T., Li, B., Lyn Santos-Cortez, R. P., Peng, B. & Leal, S. M. (2014), 'Power analysis and sample size estimation for sequence-based association studies', *Bioinformatics* **30**(16), 2377–2378.
URL: *http://dx.doi.org/10.1093/bioinformatics/btu296*

Wang, G. T., Peng, B. & Leal, S. M. (2014), 'Variant association tools for quality control and analysis of large-scale sequence and genotyping array data', *Am J Hum Genet* **94**(5), 770–83.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24791902*

Wang, R., Lagakos, S. W. & Gray, R. J. (2010), 'Testing and interval estimation for two-sample survival comparisons with small sample sizes and unequal censoring', *Biostatistics* **11**(4), 676–92.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/20439258*

Wang, X. (2014), 'Firth logistic regression for rare variant association tests', *Front Genet* **5**, 187.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24995013*

Wellcome Trust Case Control Consortium, . (2007), 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature* **447**(7145), 661–78.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/17554300*

Winham, S. J., de Andrade, M. & Miller, V. M. (2015), 'Genetics of cardiovascular disease: Importance of sex and ethnicity', *Atherosclerosis* **241**(1), 219–28.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25817330*

Winham, S. J., Pirie, A., Chen, Y. A., Larson, M. C., Fogarty, Z. C., Earp, M. A., Anton-Culver, H., Bandera, E. V., Cramer, D., Doherty, J. A., Goodman, M. T., Gronwald, J., Karlan, B. Y., Kjaer, S. K., Levine, D. A., Menon, U., Ness, R. B., Pearce, C. L., Pejovic, T., Rossing, M. A., Wentzensen, N., Bean, Y. T., Bisogna, M., Brinton, L. A., Carney, M. E., Cunningham, J. M., Cybulski, C., deFazio, A., Dicks, E. M., Edwards, R. P., Gayther, S. A., Gentry-Maharaj, A., Gore, M., Iversen, E. S., Jensen, A., Johnatty, S. E., Lester, J., Lin, H. Y., Lissowska, J., Lubinski, J.,

Menkiszak, J., Modugno, F., Moysich, K. B., Orlow, I., Pike, M. C., Ramus, S. J., Song, H., Terry, K. L., Thompson, P. J., Tyrer, J. P., van den Berg, D. J., Vierkant, R. A., Vitonis, A. F., Walsh, C., Wilkens, L. R., Wu, A. H., Yang, H., Ziogas, A., Berchuck, A., Chenevix-Trench, G., Schildkraut, J. M., Permuth-Wey, J., Phelan, C. M., Pharoah, P. D., Fridley, B. L., Sellers, T. A., Goode, E. L. & Group, A. O. C. S. (2016), 'Investigation of exomic variants associated with overall survival in ovarian cancer', *Cancer Epidemiol Biomarkers Prev* **25**(3), 446–54.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26747452*

Wu, B., Pankow, J. S. & Guan, W. (2015), 'Sequence kernel association analysis of rare variant set based on the marginal regression model for binary traits', *Genet Epidemiol* **39**(6), 399–405.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/26282996*

Wu, C., Kraft, P., Stolzenberg-Solomon, R., Steplowski, E., Brotzman, M., Xu, M., Mudgal, P., Amundadottir, L., Arslan, A. A., Bueno-de Mesquita, H. B., Gross, M., Helzlsouer, K., Jacobs, E. J., Kooperberg, C., Petersen, G. M., Zheng, W., Albanes, D., Boutron-Ruault, M. C., Buring, J. E., Canzian, F., Cao, G., Duell, E. J., Elena, J. W., Gaziano, J. M., Giovannucci, E. L., Hallmans, G., Hutchinson, A., Hunter, D. J., Jenab, M., Jiang, G., Khaw, K. T., LaCroix, A., Li, Z., Mendelsohn, J. B., Panico, S., Patel, A. V., Qian, Z. R., Riboli, E., Sesso, H., Shen, H., Shu, X. O., Tjonneland, A., Tobias, G. S., Trichopoulos, D., Virtamo, J., Visvanathan, K., Wactawski-Wende, J., Wang, C., Yu, K., Zeleniuch-Jacquotte, A., Chanock, S., Hoover, R., Hartge, P., Fuchs, C. S., Lin, D. & Wolpin, B. M. (2014), 'Genome-wide association study of survival in patients with pancreatic adenocarcinoma', *Gut* **63**(1), 152–60.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23180869*

Yip, T. S., O'Doherty, C., Tan, N. C., Dibbens, L. M. & Suppiah, V. (2014), 'Scn1a variations and response to multiple antiepileptic drugs', *Pharmacogenomics J* **14**(4), 385–9.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/24342961*

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S. & Stoica, I. (2016), 'Apache spark: A unified engine for big data processing', *Commun. ACM* **59**(11), 56–65.
URL: *http://doi.acm.org/10.1145/2934664*

Zhan, X., Hu, Y., Li, B., Abecasis, G. R. & Liu, D. J. (2016), 'Rvtests: an efficient and comprehensive tool for rare variant association analysis using sequence data', *Bioinformatics* **32**(9), 1423–6.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/27153000*

Zhang, D., Zhao, L., Li, B., He, Z., Wang, G. T., Liu, D. J. & Leal, S. M. (2017), 'Seqspark: A complete analysis tool for large-scale rare variant association studies using whole-genome and exome sequence data', *The American Journal of Human Genetic* **101**(1), 115–122.
URL: *http://dx.doi.org/10.1016/j.ajhg.2017.05.017*

Zhang, Z., Li, X., Ding, X., Li, J. & Zhang, Q. (2015), 'Gpopsim: a simulation tool for whole-genome genetic data', *BMC Genet* **16**, 10.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/25652552*

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C. & Weir, B. S. (2012), 'A high-performance computing toolset for relatedness and principal component analysis of snp data', *Bioinformatics* **28**(24), 3326–8.
URL: *https://www.ncbi.nlm.nih.gov/pubmed/23060615*

Appendix A

PHACS: COVARIATE DIAGNOSTIC PLOTS

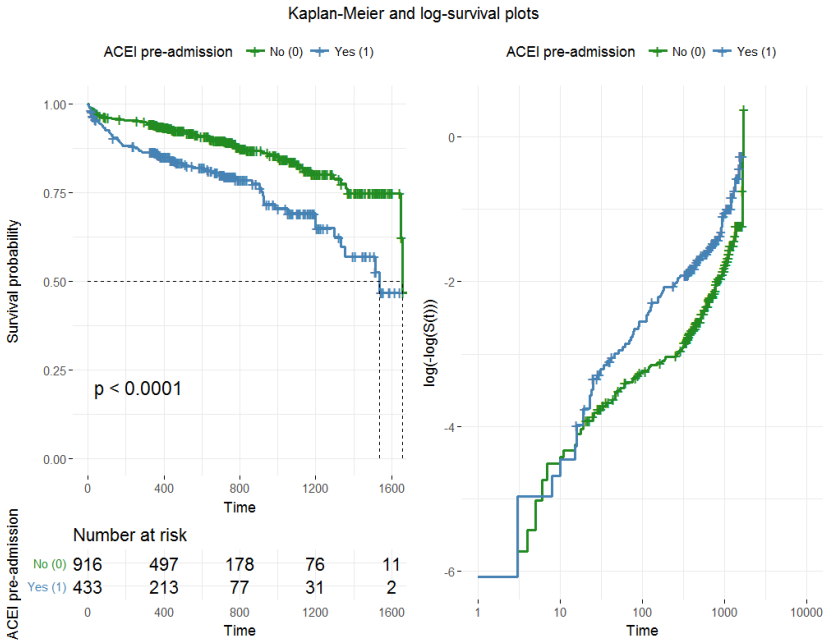


Figure A.1: ACEI: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals.

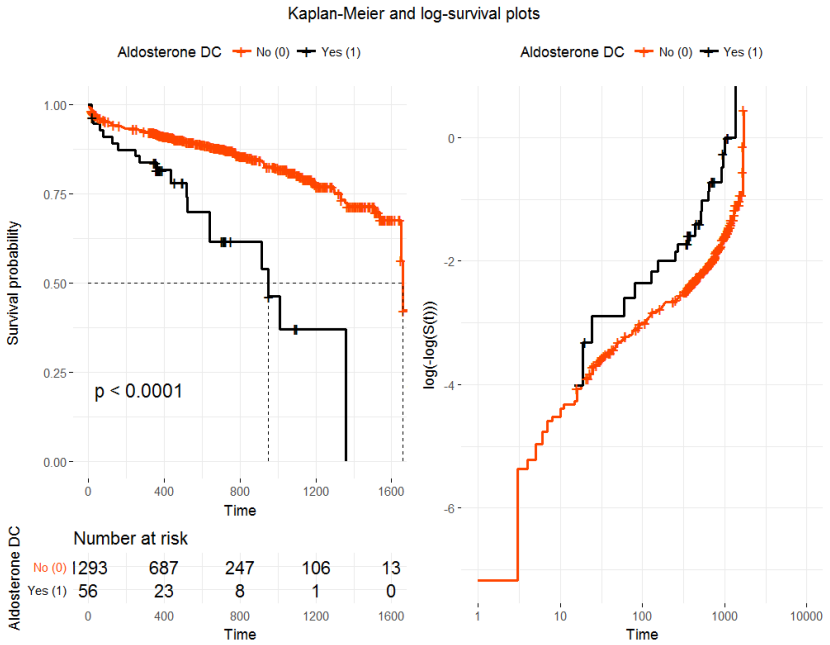


Figure A.2: Aldosterone: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals.

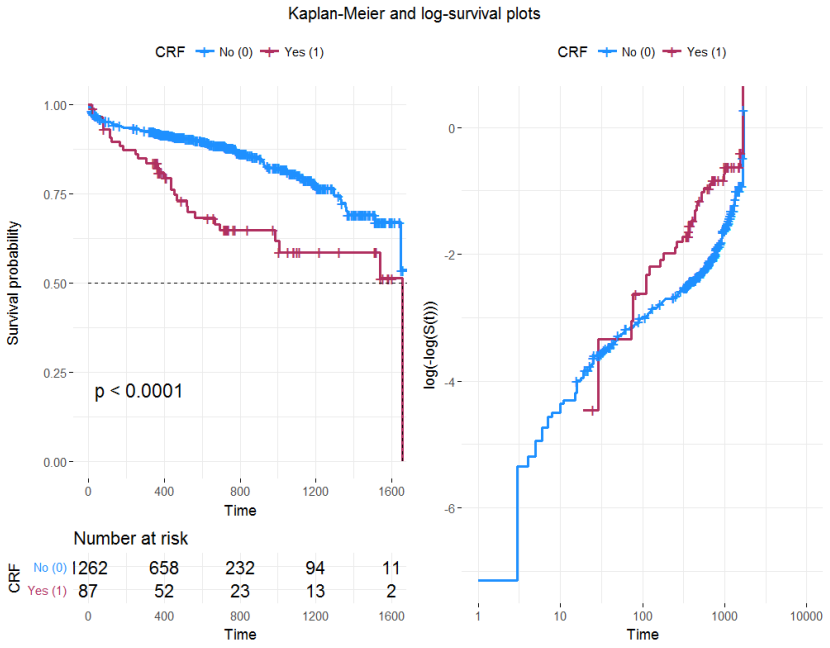


Figure A.3: Chronic renal failure: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals.

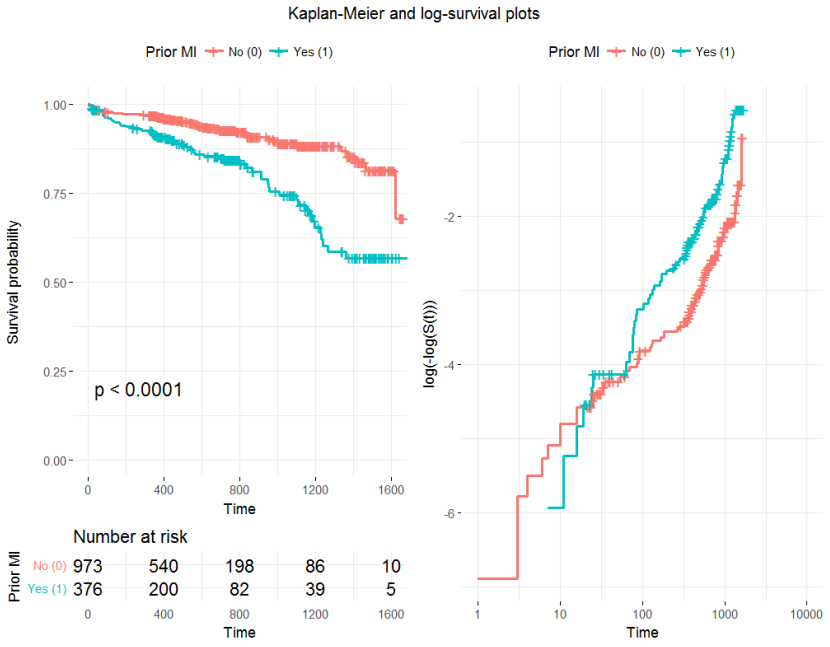


Figure A.4: Prior myocardial infarction: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals.

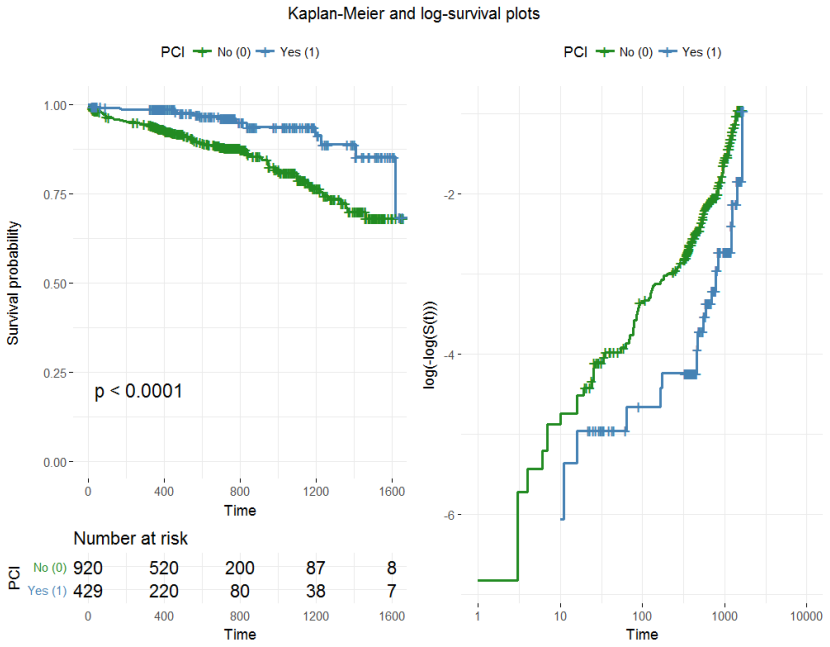


Figure A.5: PCI: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals.

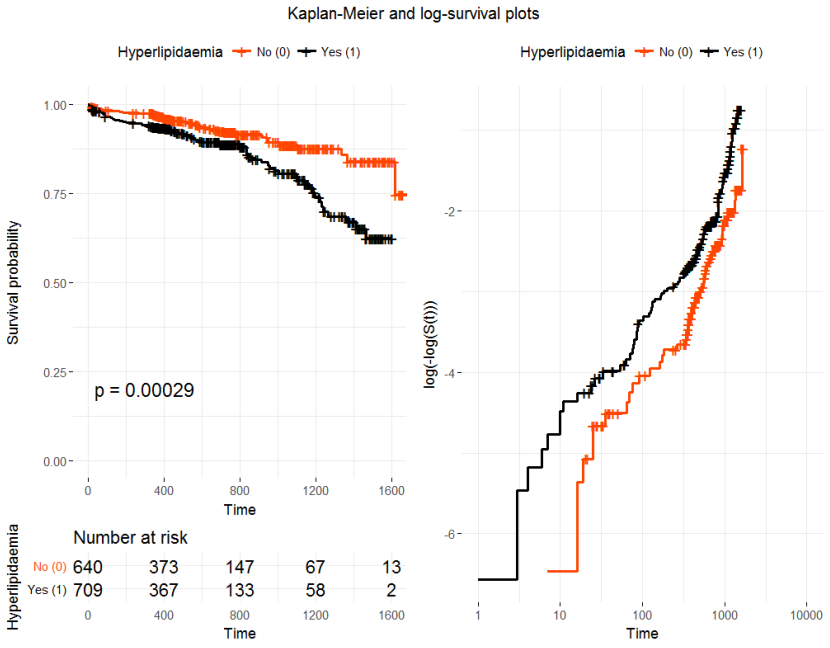


Figure A.6: Hyperlipidemia: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals.

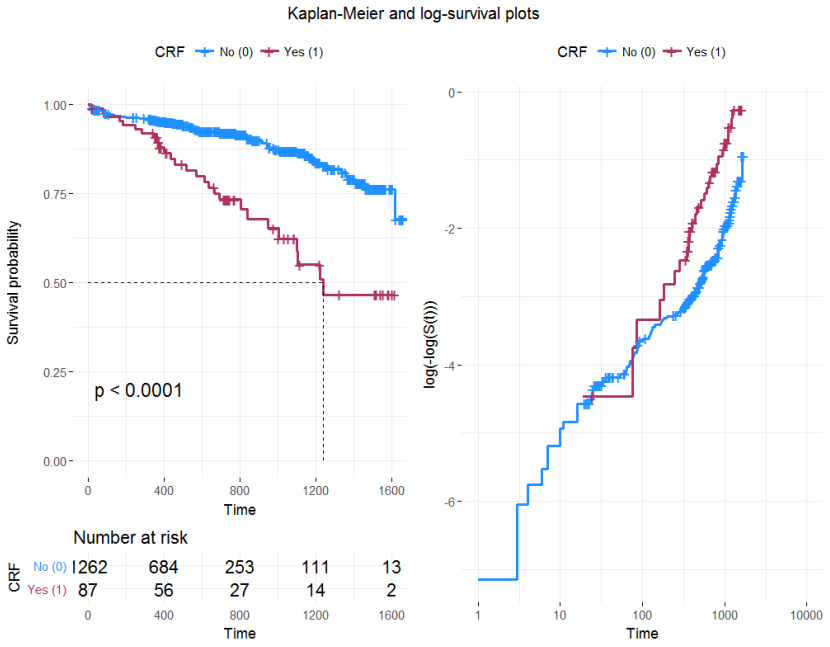


Figure A.7: Chronic renal failure: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals.

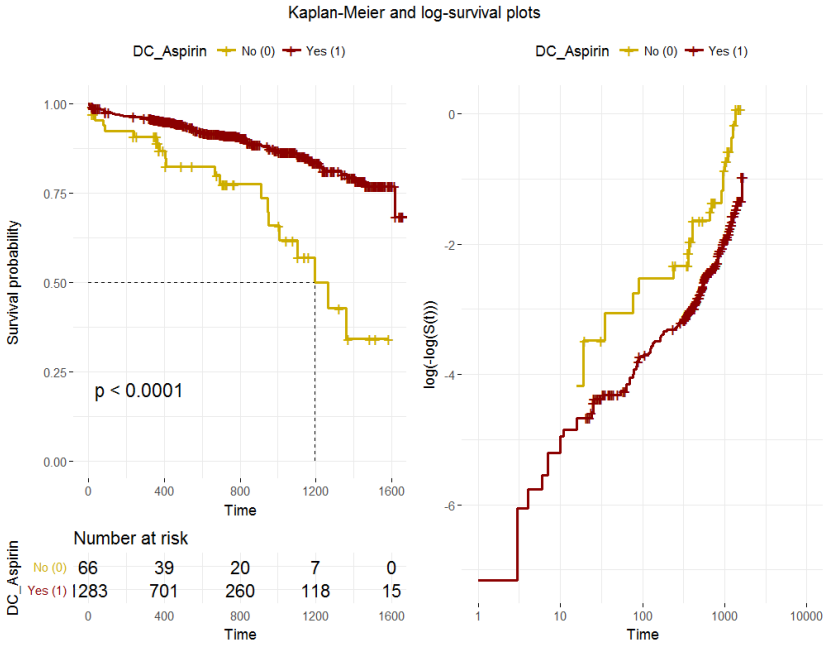


Figure A.8: Aspirin after discharge: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals.

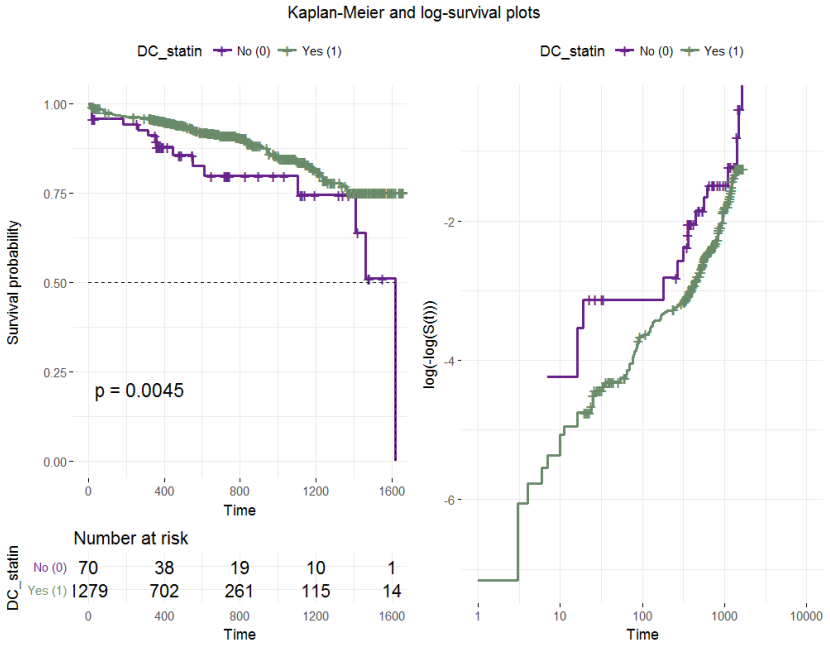


Figure A.9: Statins after discharge: Kaplan-Meier, diagnostic PH assumption plot and a summary table of at-risk individuals.

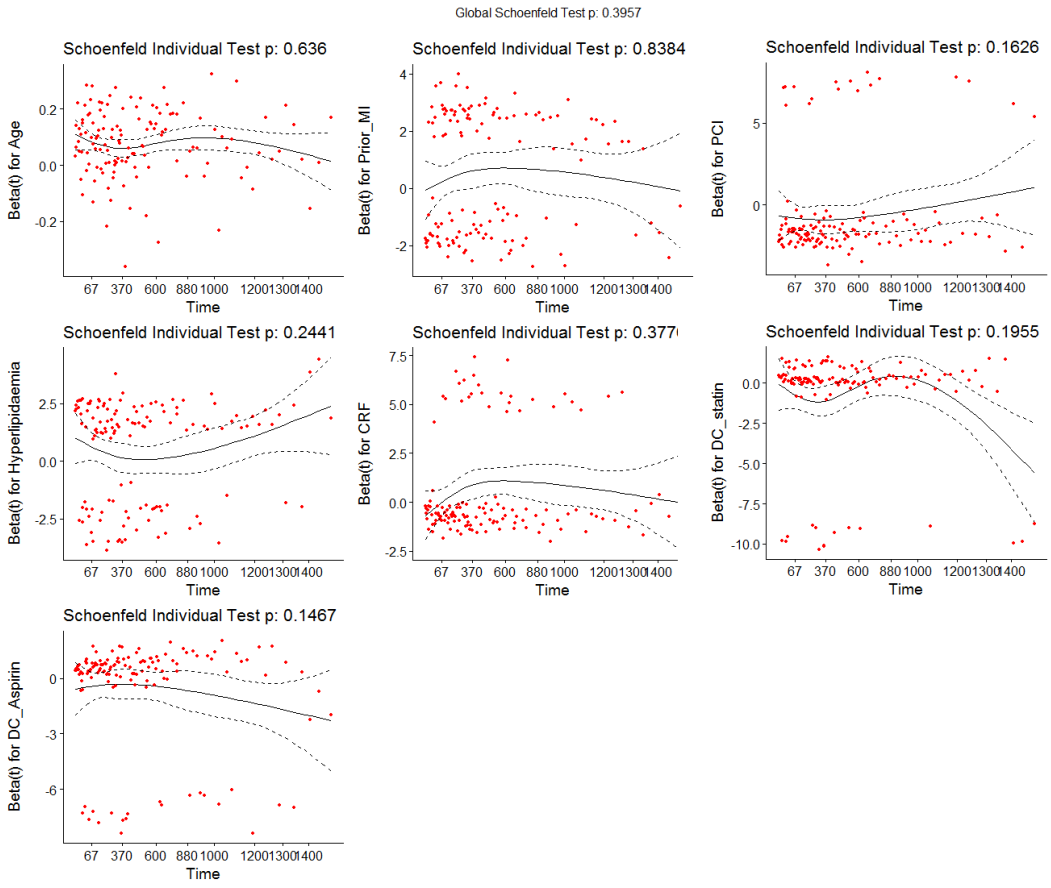


Figure A.10: Schoenfeld residual plot for each significant clinical factor with the secondary outcome.

Appendix B

PHACS: LOCUSZOOM PLOTS FOR SIGNIFICANT SNPS

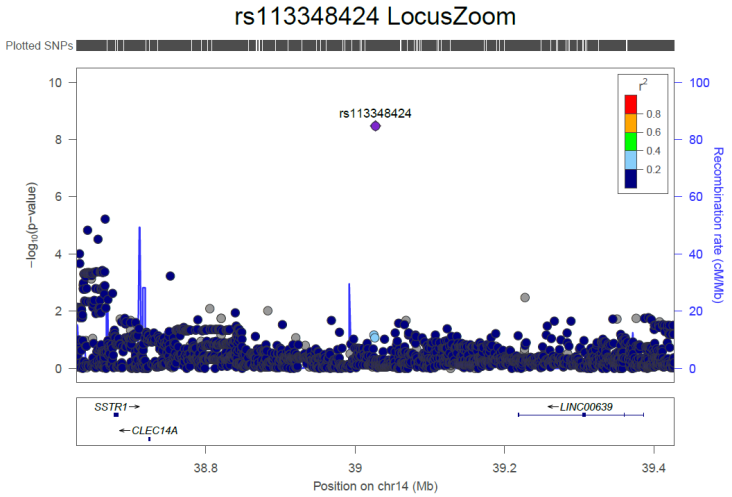


Figure B.1: Association of rs113348424 with time to a cardiovascular event. LocusZoom plot of the region associated with the primary outcome on chromosome 1 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs113348424 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs113348424 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

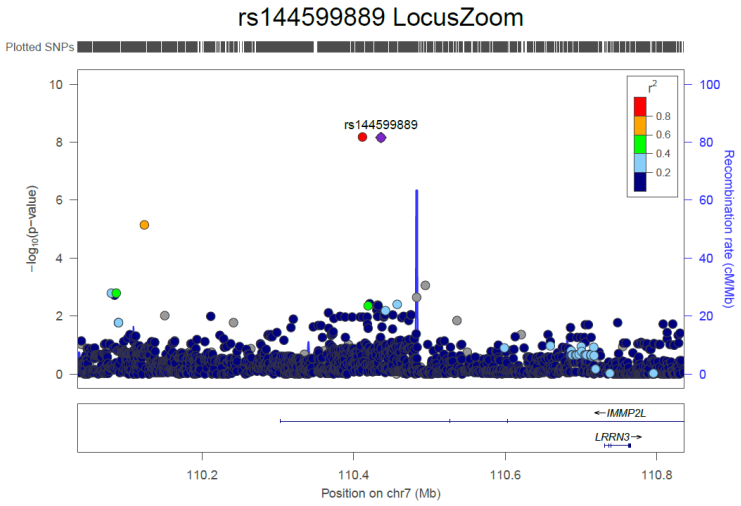


Figure B.2: Association of rs144599889 with time to a cardiovascular event. LocusZoom plot of the region associated with the primary outcome on chromosome 7 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs144599889 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs144599889 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

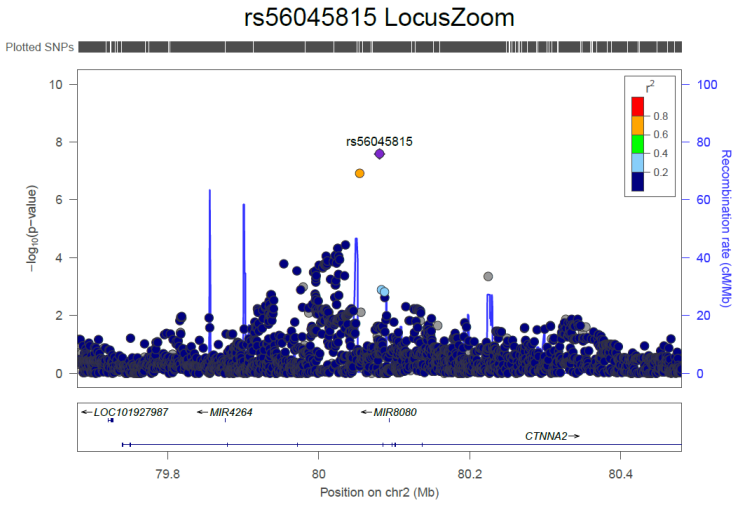


Figure B.3: Association of rs56045815 with time to a cardiovascular event. LocusZoom plot of the region associated with the primary outcome on chromosome 2 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs56045815 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs56045815 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

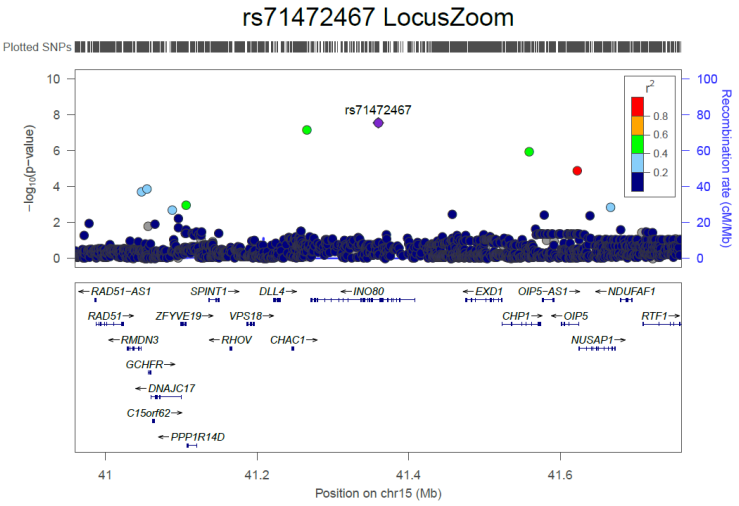


Figure B.4: Association of rs71472467 with time to a cardiovascular event. LocusZoom plot of the region associated with the primary outcome on chromosome 15 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs71472467 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs71472467 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

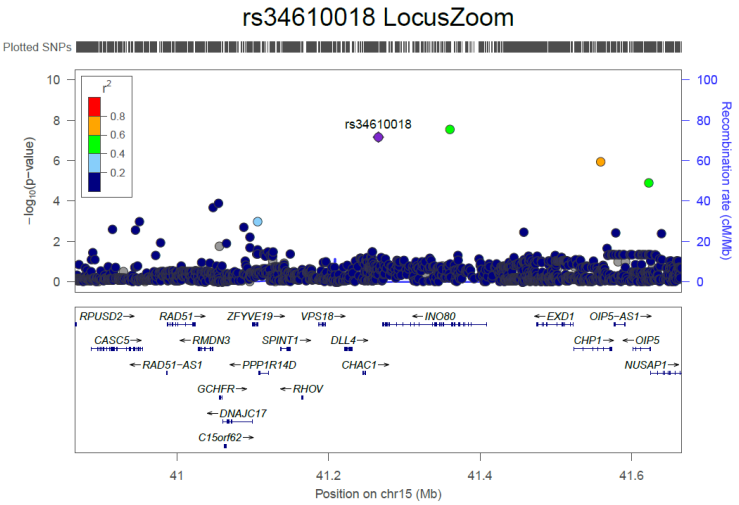


Figure B.5: Association of rs34610018 with time to a cardiovascular event. LocusZoom plot of the region associated with the primary outcome on chromosome 15 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs34610018 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs34610018 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

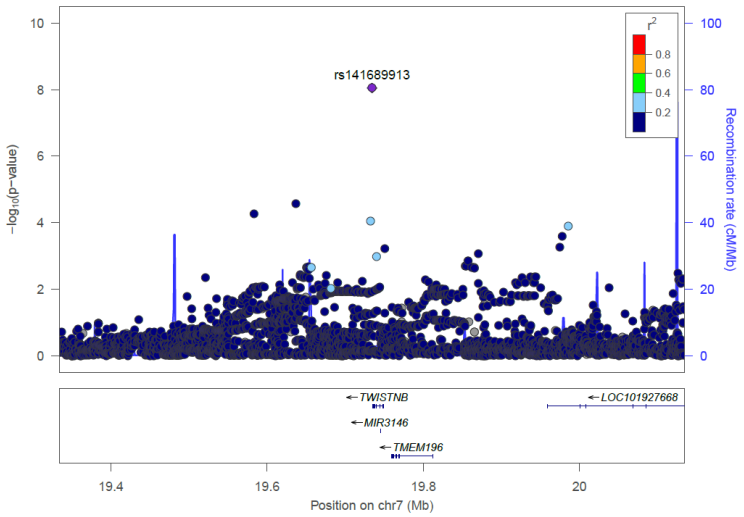


Figure B.6: Association of rs141689913 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 7 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs141689913 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs141689913 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

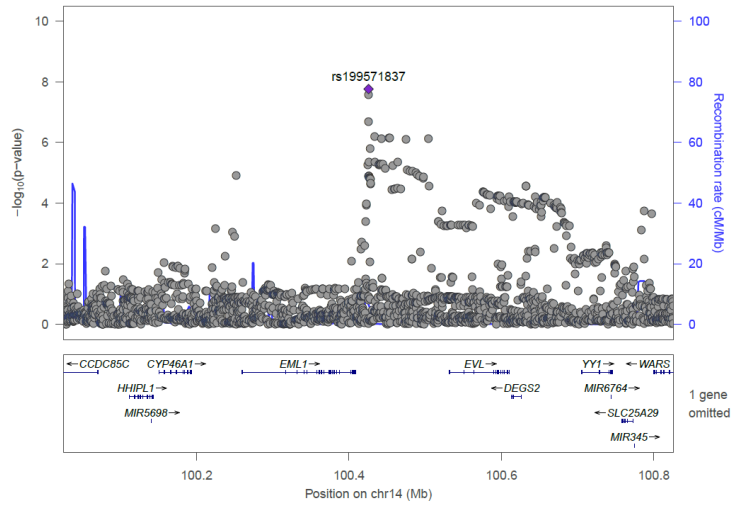


Figure B.7: Association of rs199571837 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 14 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs199571837 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs199571837 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

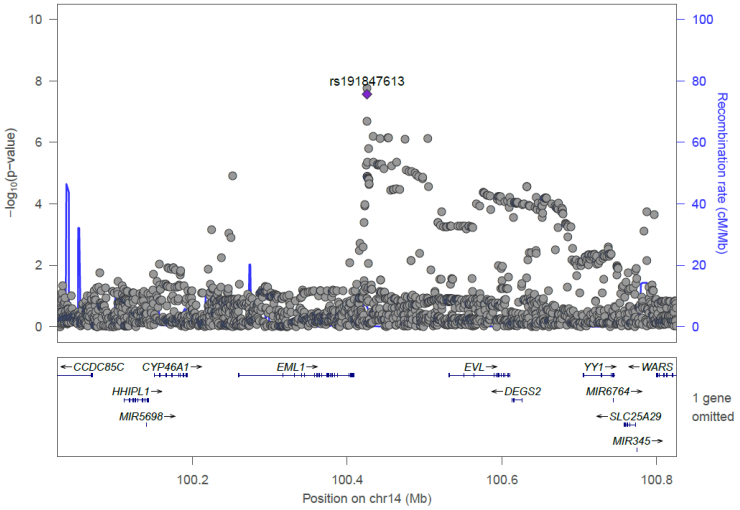


Figure B.8: Association of rs191847613 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 14 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs191847613 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs191847613 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

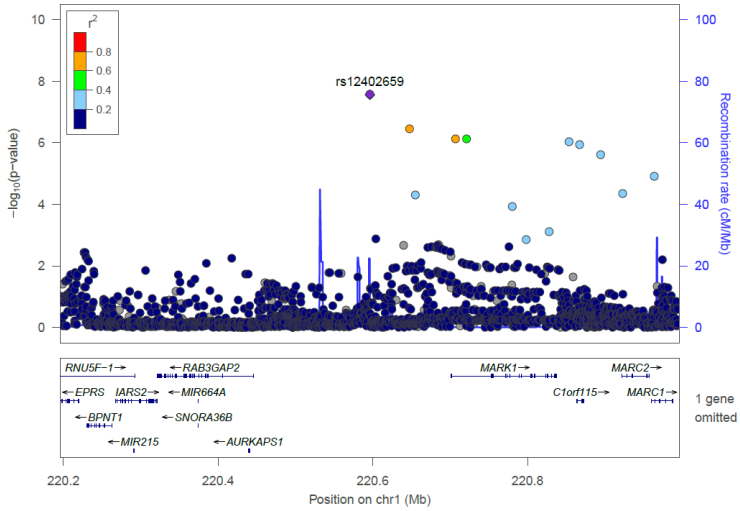


Figure B.9: Association of rs12402659 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 1 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs12402659 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs12402659 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

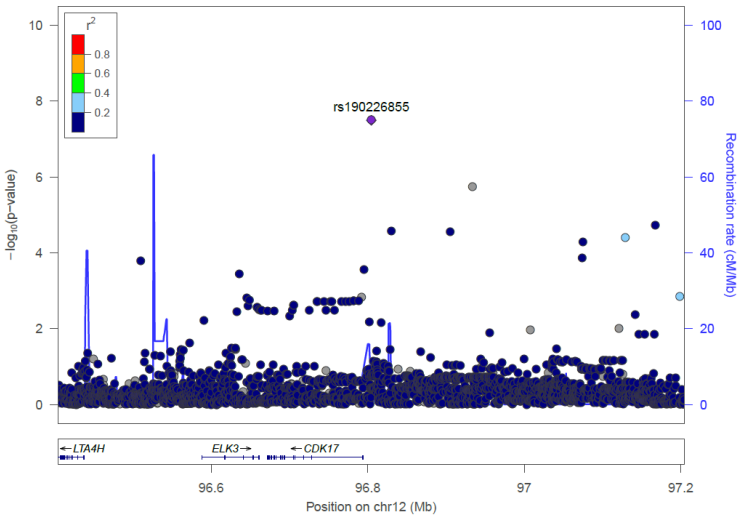


Figure B.10: Association of rs190226855 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 12 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs190226855 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs190226855 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

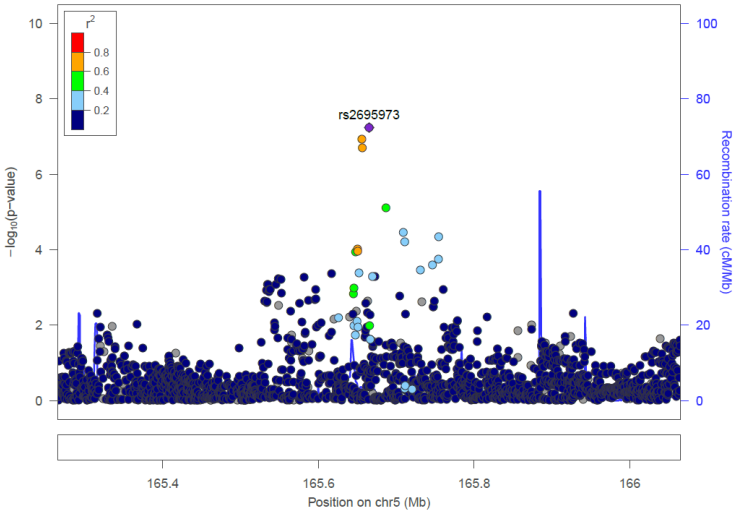


Figure B.11: Association of rs2695973 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 5 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs2695973 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs2695973 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

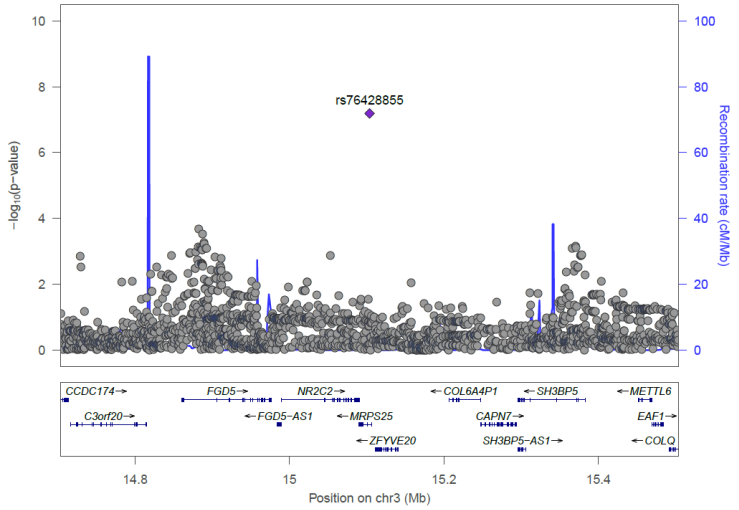


Figure B.12: Association of rs76428855 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 3 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs76428855 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs76428855 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

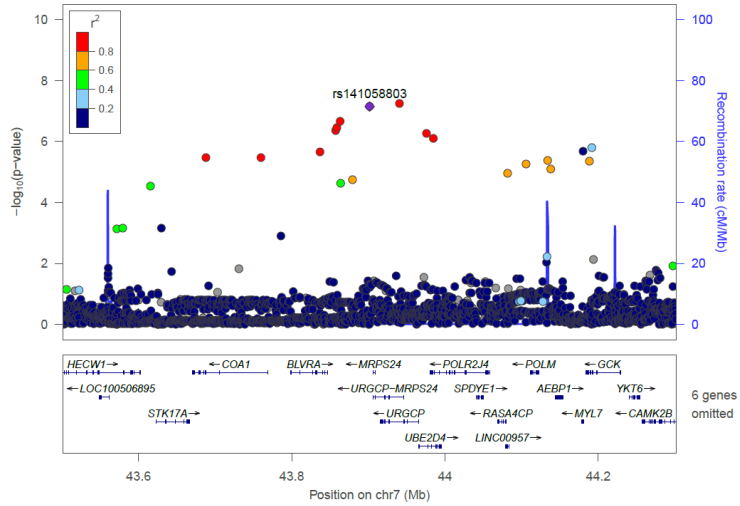


Figure B.13: Association of rs141058803 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 7 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs141058803 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs141058803 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

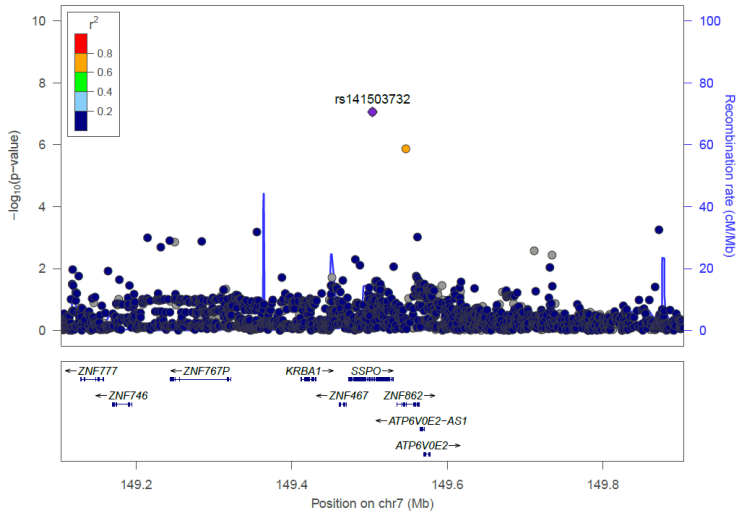


Figure B.14: Association of rs141503732 with time to all-cause mortality. LocusZoom plot of the region associated with the secondary outcome on chromosome 7 in PhACS samples. Genes within the region are shown in the lower panel, and the blue line indicates the recombination rate within the region. Each circle represents the p-value for a SNP in the discovery sample, with the top SNP rs141503732 shown in purple and the SNPs in the region coloured depending on their degree of correlation (r^2) with rs141503732 as estimated by LocusZoom from European 1000 Genomes March 2012 data.

Appendix C

PHACS: KAPLAN-MEIER PLOTS FOR SIGNIFICANT SNPS

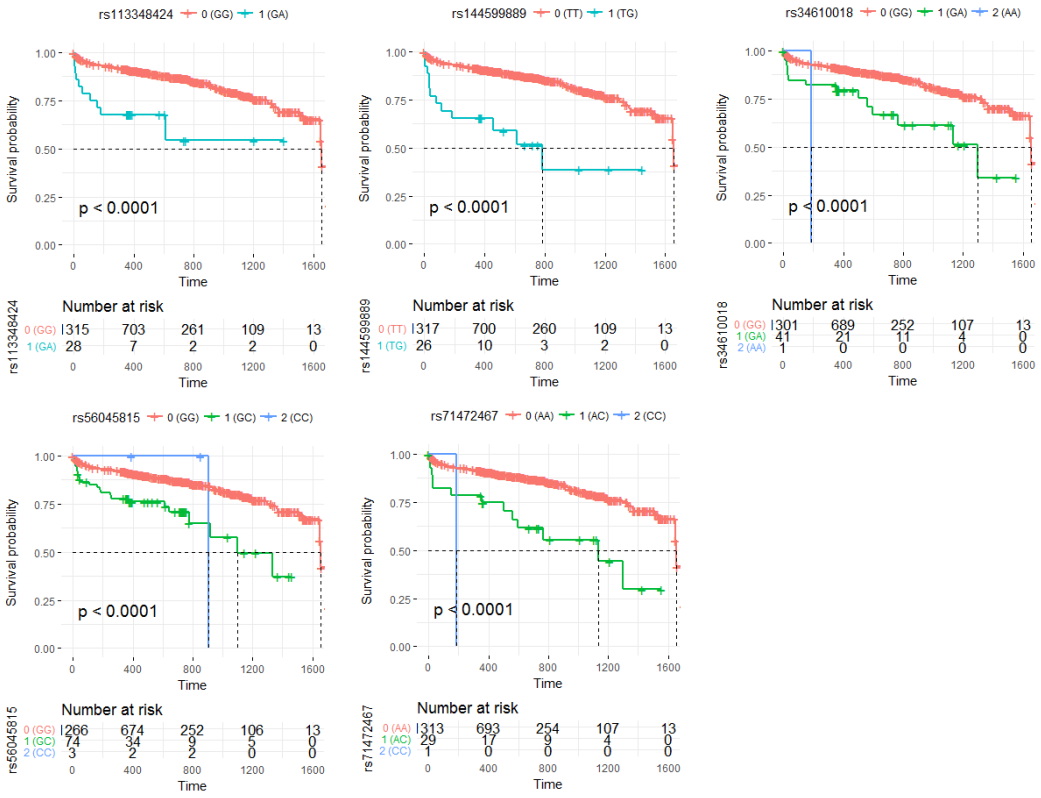


Figure C.1: Kaplan-Meier plots of genotypes for all significant SNPs associated with the primary outcome. Summary table of at-risk individuals. Top left: rs113348424, Top middle: rs144599889, Top right: rs34610018. Bottom left: rs56045815, Bottom middle: rs71472467.

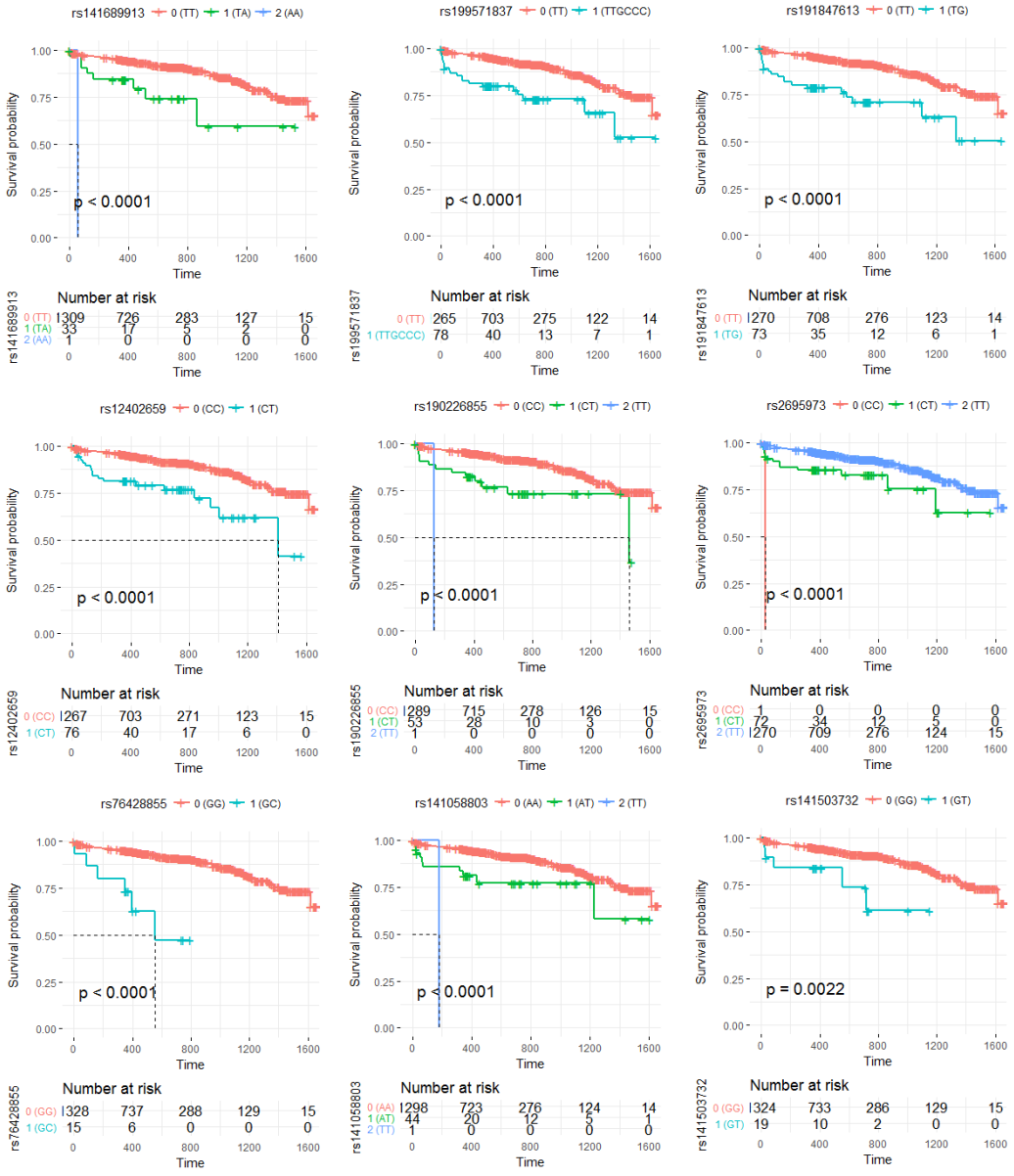


Figure C.2: Kaplan-Meier plots of genotypes for all significant SNPs associated with the secondary outcome. Summary table of at risk individuals. Top left: rs141689913, Top middle: rs199571837, Top right: rs191847613, Middle left: rs12402659, Middle : rs190226855, Middle right: rs2695973, Bottom left: rs76428855, Bottom middle: rs141058803, Bottom right: rs141503732.